notiser och
rapporter från

PEDAGOGISK-
PSYKOLOGISKA
INSTITUTIONEN

LÄRARHÖGSKOLAN
FACK, 200 45 MALMÖ 23

pedagogisk-psykologiska problem

Bierschenk, I.:

DATORBASERAD INNEHÅLLSANALYS: TEORETISKA OCH PRAKTISKA ÖVERVÄGANDEN

Nr 283 December 1975

DATORBASERAD INNEHÅLLSANALYS: TEORETISKA OCH PRAKTISKA ÖVERVÄGANDEN

Inger Bierschenk

Bierschenk, I. Datorbaserad innehållsanalys: Teoretiska och praktiska överväganden. Pedagogisk-psykologiska problem (Malmö, Sweden: School of Education), Nr 283, 1975.

Datorbaserade metoder för innehållsanalys av text har mer och mer kommit i blickpunkten inom samhällsvetenskapen. Denna rapport beskriver ett svenskt bidrag till datorbaserad innehållsanalysforskning. Metoden jämförs med internationella konventioner i fråga om behandling av text för datorbaserad bearbetning. Det redogörs för uppbyggnad av lexikon på basis av svenska språket. Kodningen av text bygger på en teori om tankestrukturer och rapporten redovisar på vilket sätt denna teori kan omsättas till ett system för kodning av begrepp och relationer. Därefter sätts regelsystemet i relation till en teori om simulering av kognitiva och emotionella strukturer. Regelsystemets kapacitet för analys på olika hierarkiska nivåer presenteras.

Nyckelord: Innehållsanalys, itemanalys, datorbaserad språkanalys, begreppsformation.

INNEHÅLL Sid

Förord

Med denna rapport vill jag beskriva ett försök med datorbaserad innehållsanalys, som är under utveckling inom projektet Skolpedagogiska sökstrategier (SÖK) vid lärarhögskolan i Malmö, och som finansieras av skolöverstyrelsen. Vetenskaplig ledare och projektledare är docent Bernhard Bierschenk. Jag själv svarar för den språkmetodiska andelen inom detta projekt.

Analystekniken, kallad ANACONDA, ställs här i en internationell belysning, dels vad gäller datorbearbetning av text överhuvudtaget, dels och främst vad gäller teorier, lingvistiska och psykologiska, som är aktuella för oss. Efter att våra första försök med denna teknik har utprövats och redovisats, har det känts viktigt att förankra tekniken i sitt ämnesspecifika och metodiska sammanhang och att visa på vilka punkter den kan jämföras med och hävda sig bland andra liknande försök. Det bör dock poängteras att vår metod för datorbaserad innehållsanalys har att passera flera stadier innan den kan betraktas som färdig. Men det är min förhoppning att den genom denna rapport har fått fastare ramar och en solidare grund än hittills att stå på inför de återstående utvecklingsfaserna.

Malmö, december, 1975

Inger Bierschenk

1. OM BRUKET AV TERMER

Denna rapport presenterar och diskuterar några modeller och teorier från såväl språk- som samhällsvetenskapliga forskningsområden. Den terminologi som utvecklats inom de olika forskningsgrenarna kan ibland försvåra förståelsen,när teorier sätts i relation till varandra. Ofta benämns fenomen olika, även om det rör sig om samma sak. Det kan också vara så, att begrepp definieras lika men används på olika sätt. Dessutom finns det termer som överlappar mellan teorierna. Ett par terminologiska problem skall exemplifieras.

I kapitel 5 används termen "conceptualization", som ibland översatts med ett försvenskat "conceptualisering", som egentligen inte säger mer än den engelska termen. Där det har känts nödvändigt har begreppet översatts med "sats", som till stor del täcker vad författaren har avsett och som är enklare att förstå, när det sätts i samband med en syntaktisk struktur.

I kapitel 6 används termen "tema" på ett helt annat sätt än vad denna term avser i de första kapitlen. Där termen nämns i kapitel 3-5 betyder den övergripande satstema eller modifiering, som tillhör en syntaktisk-semantisk språkanalys. I kapitel 6 betyder tema ungefär värderingsstruktur och tillhör en kognitionspsykologisk analys ("discourse model").

När man vill jämföra olika teorier med varandra,skulle man vilja ha en enhetlig terminologi, som väsentligt skulle underlätta jämförelsen. Men det skulle inte vara rättvist mot respektive författare eller mot teorin om man försökte anpassa termer och begrepp för ett speciellt syfte. Jag har därför valt att presentera modeller och teorier i de termer som används av deras upphovsmän och hoppas att texten ger det nödvändiga sammanhanget, så att läsaren inte känner mångfalden av de använda termerna alltför betungande för förståelsen. Däremot är det möjligt att de flesta läsare av den här skriften inte känner sig förtrogna med termer som är <u>allmänt</u> brukliga i språkvetenskapliga texter. För dessa läsare har jag sammanställt en förteckning över sådana termer med förklaringar i bilaga.

2. DATORBASERAD INNEHÅLLSANALYS: EN KORTFATTAD ÖVERBLICK

I den samhällsvetenskapliga forskningen används i stor utsträckning innehållsanalys när man vill ha information ur verbala data, t ex intervjusvar, politiska tal eller tidningstexter. Det finns emellertid ingen innehållsanalysmetod som gäller generellt för allt verbalt material. Eftersom forskare inom olika discipliner sedan länge har varit bundna till manuella analyser, har det utvecklats lika många innehållsanalystekniker som det funnits forskare som använt innehållsanalysmetoder. Först nu tycks utvecklingen ha kommit så långt att denna metod kan bli mera enhetlig och därför mera generell i och med att datorer kan användas.

Utvecklingen av en datorbaserad innehållsanalysmetod som lämpar sig för varierande syften inom de beteendevetenskapliga ämnena är i hög grad önskvärd. Den beteendevetenskapliga forskningen styrs vid informationsinsamlingen i alltför stor utsträckning av formaliserade frågeformulär. De data som på så sätt insamlats är relativt lätta att hantera jämfört med verbaliserade svar, men metoden innebär i många fall att risken för ofullständig information är stor. De forskare som inte vill ta sådana risker och därför tillåter öppna svar står genast inför problemet att bearbeta text, dvs att analysera data med i princip obegränsade tolkningsmöjligheter. Ett välkänt faktum är också att en manuell bearbetning av text är mycket tidsödande och arbetskrävande, eftersom den information man är ute efter sällan finns väl samlad. Vid organisering och strukturering av sådan information torde datorn utgöra ett viktigt hjälpmedel. En datorbaserad innehållsanalys är nästan oundviklig i de fall där forskare önskar utföra sofistikerade analyser och där de vill pröva olika teorier och modeller på samma grundmaterial utan att för den skull behöva manuellt bygga om komplexa kategorisystem och omkoda stora mängder av informationsenheter. I ett förord till "The General Inquirer" skriver sålunda Lasswell (Stone et al., 1966):

"The essential problem of content analysis is to provide a procedure that preserves and codes as economically as possible the meanings relevant to social theories. The scientist who works with samples of data with which he is concerned can now hope eventually to identify the universe of referents pertinent to his theory."

2.1 Ett system för textanalys: "The General Inquirer"

Det välkända innehållsanalyssystem som går under namnet "The General Inquirer" (GI) är en samling datorprogram, som utvecklats av P. J. Stone vid Harvard University för att användas vid innehållsanalys av naturligt språk. Stone (1966, s 5) definierar innehållsanalys som:

"any technique for making inferences by objectively and systematically identifying specific characteristics within text."

"Specificerade karakteristika" kan vara en lista på ord eller fraser, som identifieras och räknas, kategorier som bildas på lingvistisk grund, utifrån samhällsvetenskapliga teorier, etc.

GI presenterades för första gången 1962 (Harvard-versionen) och har sedan dess förändrats och förbättrats åtskilliga gånger innan den 1974 presenterades som färdig. De program som nu föreligger är till stor del ett verk av forskare i Edinburgh.

Edinburgh-versionen är i första hand en implementering av den s k Inquirer III Content Analysis på IBM 370/155. Det programspråk som använts vid Harvard är PL/1 (utvecklat av Spence), vilket inte tidigare använts i England. För det andra har man koncentrerat sig på de syntaktiska och semantiska aspekterna i Harvard IV Psycho-Sociological Dictionary, som är det huvudsakliga lexikonet vid innehållsanalyser baserade på GI (Coxon & Chalmers, 1974, ss 53-54).

GI är ett lexikonbaserat innehållsanalyssystem. Ett lexikon (dictionary) innehåller semantiska kategorier som i förväg bestämts enligt någon samhällsvetenskaplig teori. Flera lexikon har utvecklats för att ta hänsyn till olika discipliners eller problemområdens karakteristika, t ex "The Harvard Third Psychological Dictionary", "Need Achievement Dictionary" och "Psycho-Sociological Dictionary".

I en översikt över Inquirer-systemet (Trappes-Lomax, 1974, s 1) ges systemets kapacitet i följande tre punkter:

1. fysisk igenkänning av items i text;
2. igenkänning av sådana items såsom "fall" (tokens) av en specificerad lista av ord- och interpunktionstyper;
3. tilldelning av sådana items till en eller flera kategorier i ett i förväg specificerat kategorisystem.

Den första av dessa funktioner innebär ett särskiljande av ord, datum, siffror, penningsummor, vanliga förkortningar och interpunktioner (inklusive början och slut av meningar). Den andra innebär förmåga att urskilja vanliga homografer, vilket innefattar en undersökning av ordet i fråga i relation till omgivande text, interpunktion, suffixbildning och den breda klassificeringen av orden i dess omedelbara närhet. (Det skall alltså gå att komma fram till om ordet t ex är ett verb eller ett substantiv.) Därefter bestäms en av flera möjliga betydelser ("senses"). Den tredje inkluderar räkning av alla items ("tag tally procedure") för varje kategori och alla meningar som innehåller en viss kategori. De karakteristika i

texten som Stone nämner och som en analys baseras på är värdet ("value") av enskilda text items i relation till något kategorisystem i ett lexikon.

GI opererar på hel text, t ex ett helt dokument, vilket framgår av punkterna 1-2 ovan, men analysenheten är det enskilda ordet. Efter att datorn har känt igen alla ord i en text, skriver den ut en lista ("out-put file") med ord och kategoritillhörighet tillsammans med ett värde (t ex värdet från en Osgood-skala multiplicerat med frekvens i texten). Denna lista är en representation av data. Men en representation av data blir först av betydelse när den kan sättas i relation till en teori. Problemet för informationssökaren eller användaren av systemet är hur tolkningen av denna representation skall ske. Användaren måste själv ha en hypotes om vilka specifika items eller kopplingar av items som är relevanta i hans fall. I hans teori kan t ex ett ords tillhörighet till en kategori vara irrelevant. Det är ofta i tolkningsskedet som användaren finner att ett lexikon är för stelt för att vara till nytta.

En av Edinburgh-versionens huvudsakliga förbättringar av GI är utvecklingen av en procedur som markerar alla tvetydiga ord i en text och skriver ut dem med identifikationsuppgifter, så att användaren själv kan bestämma vilken kategori som i det speciella fallet är lämpligast. Denna "New Tagger-version" presenterades 1974. Om förbättringen skriver Trappes-Lomax (1974, s 4):

"Standardisation and isolation of the disambiguation operation allows the user to concentrate on modifying, as he pleases, and using more selectively, the content dictionary."

Den nya identifikationsrutinen tycks vara tillkommen efter erfarenheter av att användare nästan aldrig har teorier och hypoteser, som är så standardiserade att ett a priori-konstruerat lexikon helt passar. Visserligen är GI välkänt och mycket använt, men metoder och tekniker förändras, likaså forskningsproblem. GI är en föregångare inom datorbaserad innehållsanalys och berörda forskare har bidragit med omistliga insatser inom programutvecklingen. Det är dock tveksamt om de lexikon man inom ramen för detta system har byggt upp kan bestå, eftersom de, trots förbättringarna av forskarteamet i Edinburgh, i ringa grad tillåter nya teorier inom de samhällsvetenskapliga ämnena. En allvarlig anmärkning mot dem dessutom är att de är konstruktioner utan empiriskt underlag. Forskare som har erfarenheter av GI har vid ett internationellt forum diskuterat detta och andra system. Denna konferens, som redogörs för i det följande, har visat hur beroende ett innehållsanalyssystem är av arten av problem och hur olikartat forskare världen över behandlar sina data. För SÖK-projektets del har det varit väsentligt att ta del av dessa erfarenheter och jämföra

våra och andras framsteg. Vår analysteknik kommer därför att presenteras längre fram i rapporten i jämförande syfte. Som grund för jämförelsen presenteras först några aktuella forskningsområden och analysproblem, såsom de framförts i en preliminär konferensrapport (ISSC-CISS, 1974).

2. 2 Några problemområden inom datorbaserad innehållsanalys: Summering av en internationell workshop

Datorbaserad innehållsanalys börjar bli mer och mer använt och användbart internationellt. Kravet på program och tekniker anpassade för olika problemområden gör sig mer och mer gällande, vilket en workshop on "content analysis in the social sciences", hållen i Pisa i september 1974 har vittnat om. Denna workshop har genom sina i förväg publicerade papers (ISSC-CISS, 1974) givit en god överblick över var särskilt europeisk forskning står idag när det gäller datortekniker vid innehållsanalys. Det har känts särskilt värdefullt för SÖK-projektet att ta del av dessa papers, eftersom de flesta av dem innehåller redogörelser och funderingar kring mindre försök som fortfarande är på utvecklingsstadiet.

Redogörelsen nedan sker med reservation för ny information som kan komma att ges i en kommande bok om workshopen, redigerad av S. Draghi någon gång under hösten 1975. SÖK-projektet är representerat i papersamlingen (B. Bierschenk, 1974 a) men refereras inte här, utan mer detaljerat längre fram. De flesta bidragen är mycket knapphändiga. Framför allt skall nämnas att de sällan redogör för kodning av text eller statistisk behandling. Detta beror möjligen på att respektive författare haft för avsikt att vänta med en sådan presentation till själva konferensdagarna eller också har undersökningen eller projektet inte kommit så långt att det varit möjligt.

Konferensen i Pisa anordnades i samband med att Edinburgh-versionen av GI förelåg färdig. Avsikten var (1) att presentera den nya versionen och låta konferensdeltagarna utföra analyser med hjälp av detta system på egna data, (2) att få en samlad bild av vilka analysproblem som finns inom de samhällsvetenskapliga disciplinerna i olika länder, (3) att få belyst på vilken metodisk nivå de presenterade forskningsprojekten befinner sig (steg inom systemutvecklingen) samt (4) att försöka utforma generella riktlinjer (konventioner) för datorbaserade system, så att de får så stor räckvidd som möjligt, bl a att de kan implementeras på olika typer av maskiner.

Presentationen följer i möjligaste mån punkterna 2 och 3. Punkt 4 behandlas under egen rubrik.

Inom det politiska området ("political science") tycks ett stort intresse finnas för undersökningar där datorer kan användas vid analyser. Deichsel från Hamburg rapporterar om ett projekt som använder tidningsrubriker för att undersöka politikers image före och efter regimskiftet Brandt-Scheel 1969. Tidningsrubriker antas återspegla "political verbal interaction". Deichsel (1974, s 73) vill utveckla ett instrument för observation av den sociala strukturen av "symbol-worlds (headlines)". Genom att göra ett urval av olika slag av nyhetsorgan hoppas man kunna påvisa t ex vilken image politikerna får genom det sätt på vilket de i rubrikerna presenteras för läsarna.

Kodning och statistisk behandling av data framgår inte.

Vid Hamburgs universitet har man utvecklat EVA-systemet, baserat på GI. (EVA = Electronic Verbal Analysis.) Holzschenk (1974, s 288) redogör kortfattat för EVA-systemets kapacitet:

1. producerar listor med ordfrekvenser (med referens till text)
2. kontext-fri kodning av text på basis av ordform, utan ändelsereduktion
3. varje enhet ("tag") kan återföras till kontext ("tag-based retrieval")

Syftet med Holzschencks analys (Action Analysis) är att undersöka beroenden mellan den handlande och handlingen i tidningsrubriker. Tyska politiska partier är de handlande ("actors") och objekt för handlingar ("objects of action").

Ett lexikon har byggts upp, kallat SEMANTAX, som innehåller begrepp (concepts) av första eller andra graden, således huvud- och underavdelningar, av vilka de senare används efter en validering genom textsökning. Dessutom finns syntaktiska markeringar.

Typen av analys beskrivs på följande sätt (Holzschenck, 1974, s 289):

1. kodning av text i sekvenser
2. "tag"-baserad kontextsökning
3. syntaktisk/semantisk rekonstruktion av meningar

Dataanalysen innebär bildandet av frekvensordlista av ord eller begrepp ("tags") i definierade syntaktiska positioner. De resultat man väntar sig att finna är en struktur i politiska partiers handlingar som "actors" och "objects of action".

Andra europeiska forskare med politiskt intresse är Susan Garbett i London, Karen Siune i Aarhus och Ulf Himmelstrand i Uppsala. Garbett vill undersöka interaktionen mellan politikerna i parlamentet och massmedia, framför allt TV. Problemställningen är: Vill parlamentet vara med om att spelas in under sessionerna? Undersökningsmetod tycks inte

vara fa. :ställd, inte heller innehåll analysteknikens art. Lika vag i detta avseende är Siunes framställning. Hon är intresserad av valkampanjen 1971 och hur kommunikationen mellan partierna och väljarna sker via radio och TV. Den teori som ligger bakom är frågan om huruvida valpropaganda är "plattformar" för politikerna eller "fönster" för väljarna. Siune presenterar siffror för olika variabler, som t ex art av framträdande (utfrågning i TV, debatt-tal) eller slaget av ämnesval (ekonomi, familjepolitik). Till vad dessa variabler skall användas tycks inte vara klarlagt ännu, vilket också medför att man inte får någon anvisning om hur en analys via dator skall kunna ske.

Himmelstrand, tillsammans med bl a K-E Rosengren i Lund, tycks ha kommit mycket längre. Detta forskarteam intresserar sig för hur kriget i Nigeria kommunicerats i internationell press. Problemställningen är inte så olik forskarnas i Hamburg. Politisk "actor" eller "event" står i förgrunden. Problemställningen är av komplex natur och syns i detta paper vara väl underbyggd. Sammanfattningsvis kan man kalla undersökningen en studie i relationerna mellan samhälls- och organisationsstruktur och dessa strukturers manifestering i massmedia. Syftet är att undersöka om bakgrundsinformation av strukturell natur (t ex karakteristika hos en politisk "actor" eller "event") korrelerar med tidigare använda moraliska omdömen i politiska kommentarer till rapporterade händelser. Hypotesen är att starka moraliska omdömen uppträder även där det finns rik kontextuell information av strukturell natur i rapporteringen, men att sådana värderande omdömen uttrycks på ett senare stadium, eller om de uttryckts tidigare, är omgivna av reservationer (Himmelstrand et al., 1974, s 42). Många variabler är av intresse i denna omfattande studie, t ex afrikanska nyhetsorgans interna struktur, nyhetsförmedlingen inom Afrika, m m.

Undersökningar i form av pilotstudier rapporteras ha företagits. Ett omfattande kodat material finns. Men på vilket sätt det skall datorbearbetas framgår tyvärr inte.

Inte enbart massmedia är objekt för politiska studier. Vid universitetet i Giessen har Mohler försökt lägga upp en undersökning av "individuell och kollektiv makt". Stoffet hämtar han ur tyska pojkars examensuppsatser från 1917-1971. Mohler (1974, s 300) säger:

"My problem is to extract political images and to show their change over time."

Om image-strukturen förändrats genom tiden förväntar författaren att det återspeglas i antalet klassificerade teman ("emotion", "action," m fl). Han efterlyser dock ett medel att återvinna teman i en text, eftersom en "one-man-analysis" som denna skulle vara mycket tidsödande. Ett "general dictionary" vore önskvärt.

Namenwirth (Connecticut) har studerat förändringar i amerikanska värde..ingsmönster (American Party Platforms). Det har funnits olika slag av värderingar under det sista århundradet, t ex expressiva, adaptiva, instrumentella, osv. På vad sätt värdeförändring sker mäts med klusteranalyser. Slutsatsen är att ju större konflikter ett samhälle har, desto större "value intensity" finns det (Namenwirth, 1974, s 168).

I beteendevetenskapligt sammanhang finns det också ett stort intresse att genom analys av naturligt språk via datorbearbetningar komma åt en struktur eller mening.

Vid universitetet i Ulm undersöker en grupp (Kächele et al., 1974, ss 100-110) förändringar i innehållet i tal under psykoanalytisk behandling. Variablerna är substantiv, som tilldelas semantiska kategorier. Genom statistisk analys (korrelationsstudier) får man fram vilka begrepp som väcker eller uttrycker negativa resp positiva känslor hos patienterna. Kliniska variabler, t ex ångestneuros, kan således definieras med hjälp av verbalt innehåll, vilket konfirmerar de antaganden man har i psykoanalytisk teori om neuroser. Egna program och lexikon har inte byggts upp, utan man använder sig av GI.

Vid Zentralarchiv für Empirische Sozialforschung, universitetet i Köln, är Kops intresserad av äktenskapsannonser. Han vill studera dem ur en vad han kallar "Who-am-I"-vinkel. Annonserna beskriver den egna personen och vilka egenskaper man förväntar sig hos den sökta parten. En annons som enhet är alltså en self-other-beskrivning. Genom att välja tidningar som man vet läses av olika socialgrupper (Die Zeit - Heim und Welt) förväntar Kops sig att finna skillnader i socialgruppernas normer och värderingar. Undersökningen har inte kommit så långt, då man fortfarande (1974) har problem med kodning och stansning. Däremot har Kops tillgång till ett väl utvecklat datoranalysprogram i Köln, som kallas TEXTPACK och har utvecklats vid denna institution.

Tiemann (Freie Universität, Berlin) anmäler problem med lexikonkonstruktion angående "Ingroup-Outgroup-Analysis". Han diskuterar problem med att skapa kategorier som inte överlappar, vilket GI-systemet gör. Han har just de analysproblem som nämndes i samband med Edinburgh-versionens förbättring för användaren (se sid 6).

Transgaard från Köpenhamns socialforskningsinstitut är intresserad av teorier och modeller för attitydstrukturer och vilka metoder man bör använda för att komma åt attityder, samt hur kodning och analys kan ske, framför allt när intervjuer används som undersökningsteknik (Transgaard, 1974, s 308).

Av mer antropologisk natur är Marandas studier av myter. Hypotesen

är att myter, som semantiska system, bildar semantiska kartor med en probabilistisk struktur och att de kan beskrivas genom användning av automata-teorin. Maranda & Maranda (1974, s 138) säger:

"... human communication is a social contract which rests on subliminal laws, and /../ a culture's myths contain its semantic jurisprudence."

Denna studie kan inordnas bland sådana som försöker analysera semantiska nätverk.

Samma författare har också studerat rollkaraktäristik ur Racines verk Andromaque (Maranda, 1974, ss 110-123). Undersökningen gäller spelet mellan personerna och förhållandet mellan dem samt sannolikheten för att den ena personens utspel följs av replik från en viss annan. Det finns dessutom ett samband mellan handlingsfortskridande och rollens klasstillhörighet. Han får fram ett handlingsschema eller nätverk som visar att huvudkaraktärerna bygger upp handling.

Information och dokumentation är ett specialområde som utvecklas mer och mer. Även här tycks Köln ha en framskjuten plats. Mochman (1974, s 286) presenterar Z.A.R.-systemet, som lagrar och återvinner data. I stället för indexering finns det utvecklat ett "retrieval dictionary". Systemet möjliggör återvinnande av samtidigt förekommande och automatiskt särskiljande av relevanta termer och har rutiner för morfologisk analys. Mochman (1974, s 298) anmäler systemet på följande sätt:

"Compared to Content Analysis, which has been defined a technique for making inferences, Information Retrieval can be labelled a technique for making references."

Programmet RIQS (Remote Information Query System) har utvecklats och förbättras fortgående av Northwestern University in USA. Systemet är i första hand tänkt att användas på bibliotek och handhar bibliografiska data. (Används bl a i Amsterdam.)

De program för viss innehållsbearbetning av text som hittills nämnts i paper-samlingen till Pisa-konferensen ger ingen presentation av hur programmen används. Ett bidrag som endast har till syfte att detaljerat ge sådana instruktioner är utarbetat av Spence (Research Center for Mental Health, New York University). Programmens kapacitet redovisas och förutsättningar för kategoriers uppbyggnad på hålkort ges. Spence visar programmen DICT, LIST, HUNT, SEARCH och SORT. Hela paketet kallas PL/1, som är programspråket för WORDS- och Inquirer-systemen.

Mindre känt är kanske det norska bidraget till datorbaserad innehållsanalys. Fonnes presenterar TEXT, "a general program package for text processing". Programmet har utvecklats för att underlätta återvinning av speciella items i stora textmängder. Det mest positiva med TEXT är att

det inte kräver ingående kunskaper i programmering, utan kan användas av humanister efter en kort introduktion (Fonnes, 1974, s 78). Fonnes nämner dock att TEXT inte i första hand var avsett för innehållsanalys, men att en utveckling av paketet pågår vid universitetet i Oslo.

De redovisade projekten och undersökningarna är endast en liten del av vad som pågår inom området datorbaserad innehållsanalys. Den mycket omfattande amerikanska satsningen är i synnerligen liten utsträckning representerad här. Det är dock i USA som pionjärinsatserna har skett. Men även i Europa satsas det på datoranalyser. Ledande land här är otvivelaktigt Västtyskland. Framför allt Köln och Hamburg har utvecklat gedigna programpaket.

Av de redovisade bidragen att döma tycks de flesta försöken beröra sociologiska undersökningsfält. En förklaring till just denna typ av bidrag till konferensen kan vara att P. Stone själv är sociolog, och att informationen om denna konferens huvudsakligen nått universitetens sociologiska institutioner. Intresset för denna teknik inom de samhällsvetenskapliga ämnena får ändå anses väl dokumenterat genom bidragen till den internationella konferensen i Pisa. Stone (1975 a, ss 110-111) säger:

"It is no longer a matter of a few pioneer systems only available at extremely large computation centers in extremely advanced countries."

men tillägger också att:

"... Pisa also made it evident that many problems remain to be solved."

Som en summering av konferensens erfarenheter beslöt man att utarbeta några konventioner och rekommendationer, som i fortsättningen borde följas av utvecklare och användare av datorer i samband med innehållsanalys. Dessa s k "Pisa conventions" redovisas nedan och följer Stones framställning (Stone, 1975 b, ss 127-137).

2.3 Konventioner för textbehandling

Stone konstaterade efter konferensen bl a (Stone, 1975 a, s 107) att "Content analysis researchers are not coordinated at all."

Innehållsanalysforskning bedrivs på ett dussintal språk och med stor variation ifråga om strategier. De program och erfarenheter samt även nyare frågeställningar och problem som konferensen samlade blev en utmärkt grund för generella diskussioner.

Den kanske viktigaste frågan att diskutera vid innehållsanalyser gäller analysenhet ("unit"). Vanligen är det det enskilda ordet som används i dessa analyser, eftersom det är lämpligast för olika slag av statistiska bearbetningar. Men ordet bör inte vara skilt från kontexten. Meningen

("sentence") är en lämplig enhet vid tema-analys (t ex förmodan). Dessutom kan det vara så att en ordsträng, dvs en räcka av ord, är en meningsenhet ("content"), t ex vid idiom.

Konvention 1. Analysenhet är det enskilda ordet. Varje enhet eller "tag" bör innehålla information som identifierar ordet till visst dokument, stycke, mening och meningsposition. När man arbetar med hel text som input är det också viktigt att man kodar interpunktionen. Detta möjliggör att man kan analysera teman i ett textavsnitt. Motiveringen för att ta ordet som bas är att olika slag av ordlistor, alfabetiska eller frekvensbyggda, är enklare att skapa än sådana som bygger på syntaktisk struktur.

Konvention 2. "Common derived text formats". Standardisering för hur text skall matas in i dator är nödvändigt. Om data skall stansas skall det vara i form av versaler med numerisk information i början eller slutet av kortet. Betydelsen av parenteser kan variera beroende på vilket format som används. Men i stället för att fixera ett format vore det värdefullt att tillåta flera slag av inmatningar, som sedan kan konverteras till ett standardformat.

Konvention 3. Information om identifikation. Beroende på text eller informationstyp bör det finnas plats för kodning av identifikationstyper som är relevanta i det speciella fallet. I en intervju t ex kan man vara intresserad av distinktionen mellan fråga och svar. Ett fält för sådan identifiering bör kunna variera i längd. Varje analysenhet innehåller således information som är textbaserad (syntaktisk eller semantisk kodning) och/eller icke textbaserad (identifikation).

Förutom att man enades om dessa tre konventioner beslöt man också att rekommendera ("Pisa recommendations") följande:

1. separering av typ och fall ("type-token") i lexikon och
2. anpassning av lexikonsökningsstrategier till respektive språk

Den första punkten tar upp frågan om hur varje fall ("token") av ett ord skall representeras i ett register. Varje ord skall följas av deskriptorer (som här betyder kategoritillhörighet och värde) och det bör räcka med ordtyp. När detta ord uppträder flera gånger i texten skall samma deskriptorer inte behöva upprepas.

Det kan emellertid uppstå problem med att definiera vad som är ett fall av samma ord, dvs man glider in på innehållet i varje ord, ifall det uppträder i olika syntaktiska positioner. Har ordet då en ny betydelse som skall registreras som särskild typ? Stone själv kommenterar detta (Stone, 1975 b, s 133):

"If so, then the type file is not stable, but in fact acquires more records as the processing continues and new distinctions are added."

Den andra punkten tar upp problem med ändelser. Om den s k "dictionary lookup"-rutinen sorterar ord alfabetiskt och sedan jämför orden med ett lexikon, uppstår besvär med alla böjningsformer om dessa inte finns i lexikonet. Stone refererar till systemet WORDS som har ett lexikon med ett stort antal böjningsformer. För varje ordböjning finns en rotform som ersätter den böjda formen. Oregelbundna former ersätts med regelbundna. Men detta system kan inte användas på t ex ett språk som tyskan, som har mångdubbelt fler ändelser än engelskan. Slutsatsen är att ett lexikon uppbyggt med alla böjda ordformer blir onödigt stort och ohanterligt. Inquirer-systemet har en s k "hash list"-sökning. Den innebär att datorn söker på ordlängd. När hela ordet har hittats tas ändelsen bort. Denna procedur fungerar när suffixen är kända och finns listade i ett särskilt ändelseregister.

Ett annat exempel på olikheter mellan språk är ordsammansättningar. Om sammansättningsleden skall betraktas som ett ord var för sig, skulle i tyskan t ex suffixlistan bli oändlig och således oanvändbar. Stone tar exemplet "Sozialforschung". Det enda sätt på vilket en procedur som ovan nämnts skulle kunna komma fram till ordet "Sozial", vore att betrakta "Forschung" som en ändelse. Anledningen till denna diskussion är att liknande sammansättningar inte finns på engelska, där orden är delade, "social research".

Pisa-konferensen har visat att hur väl utvecklade datorprogrammen än tycks vara, så lämpar de sig inte för alla språk. Innehållsanalysforskare måste därför anpassa sin programutveckling till det egna språkets karaktär. Ett system på svenska kan således utnyttja vissa principer i redan existerande system, men detaljerna måste utarbetas beroende på det svenska språkets uppbyggnad, och det är både tids- och kostnadskrävande. Detta faktum kan man inte gå förbi och det ensamt räcker till för att förklara varför "content analysis researchers are not coordinated at all".

3. ANACONDA OCH PISA - KONVENTIONERNA

SÖK-projektets metodutveckling för en datorbaserad innehållsanalys (Analysis of concepts by data processing, ANACONDA) har nu pågått i två år. Bakgrunden till projektet, där denna del ingår, har redovisats i flera tidigare sammanhang. Projektets uppbyggnad och innehåll har presenterats av B. Bierschenk (1974 b). Här lämnas ändå några upplysningar som kan vara av vikt för en jämförelse med andra liknande försök.

Vårt problemområde faller närmast inom disciplinen psykologi och rör "forskning om forskning". Forskare på beteendevetenskapliga institutioner i Sverige har intervjuats om sin forskningssituation, idéer till forskningsproblem, strategier och tekniker vid genomförandet av forskningsuppgifter samt sättet att samla information om problemområdet ifråga. Datainsamlingen har skett via intervjuer med såväl bundna som öppna svarsalternativ. Vårt försök med datorbaserad innehållsanalys har till uppgift att strukturera de öppna svaren, som vid intervjutillfället togs upp på ljudband. Våra försök med bearbetning av text har gjorts på ca 10% av detta material, som totalt omfattar nära 4000 sidor text. Det frågeområde som kommit att behandlas i detta skede gäller information och dokumentation, som betraktades som tillräckligt konkret för prövningen av metodutvecklingens olika faser.

Syftet med vår typ av innehållsanalys är att försöka få fram en sådan struktur i intervjupersonernas yttranden som återspeglar vem (agenter eller "actors") som agerar (planer och strategier) mot vem eller vad (objekt eller mål). Paradigmet förkortas A a O. För en innehållsbestämning av de enskilda items eller textkarakteristika använder vi Osgoods teknik för skalering av adjektiv och verb i de tre dimensionerna positiv-negativ, aktiv-passiv och stark-svag. De värden vi där får fram används för en kvantifiering av "attitydobjekt". Vi avser bygga upp strukturerade register av begrepp på basis av bl a klusteranalyser och på så sätt undvika att i förväg bestämma kategorier (jfr GI).

Syftet med följande presentation är att belysa likheter och olikheter mellan ANACONDA och GI, eftersom den senare ligger till grund för många andra analysförsök, vilket också Pisa-konferensen har visat.

De konventioner som Pisa-konferensen utmynnade i är viktiga spörsmål för alla som arbetar med datoranalyser för att hela området skall kunna stabiliseras och för att utbytet mellan de olika forskarna skall bli mera fruktbart. Däremot har rekommendationerna inte tillkommit för att slaviskt följas, utan för att vara underlättande för kommunikationerna mellan programutvecklarna och innehållsanalysforskarna. I vad mån fors-

karna anammar en rekommendation är beroende av vilken typ av information varje analys är tänkt att ge. Det vi i det här fallet är intresserade av är i hur stor utsträckning vi följer konventionerna, framför allt med tanke på att våra försök med inmatning och identifikation startade före konferensen. För oss är det därför viktigt att veta att vi inte ligger för långt ifrån dessa allmänna riktlinjer när vi vill föra ut systemet i internationellt sammanhang.

Framställningen nedan skall följa Pisa-konventionernas punkter och där det är nödvändigt dras en parallell till GI.

3.1 Analysenhet

En väsentlig och grundläggande skillnad mellan vår analys och de flesta andra är att vi baserar vårt kodsystem på talad text. GI använder skrivna dokument av olika slag. I vårt fall måste den talade texten först prepareras innan vidare hantering kan ske, eftersom den är nedskriven från ljudband i så "autentiskt" skick som möjligt. Det går alltså inte att låta stansa denna text i det skick den befinner sig på bandet. Först måste en rensning ske, så att den går att urskilja i meningar och satser, vilket medför att det måste finnas regler för kodare att först åstadkomma denna rensning och urskiljning (se I. Bierschenk, 1974, s 50). För att illustrera hur vårt autentiska material kan se ut ges i ruta 1 ett textavsnitt med samma text i preparerat skick.

Vår analysenhet är meningen ("sentence") och inte det enskilda ordet. Varje mening analyseras för sig i satser och de där ingående delarna, dvs begrepp, som kodas enligt funktion i satsen. Input är således begrepp, som kan vara ett enda eller flera ord. Om man stansar in hel text utan föregående segmentering återstår ett omfattande arbete med automatisk igenkänning av en texts olika delar, som vi inte har haft för avsikt att försöka utveckla, dels därför att sådana försök har gjorts till en vetenskap för sig och dels för att problemen syns oöverstigliga när det gäller de semantiska aspekterna på textförståelse. Visserligen tar kodningen av texten i vårt fall lång tid i anspråk, men när den väl är genomförd har vi ett analysinstrument med många fler möjligheter än många andra, som inte kan åstadkomma så mycket mer med sitt material än att lista ord. Och den möjligheten går ju inte förlorad för oss ändå.

Stone anförde i avsnittet om den första Pisa-konventionen att möjligheterna att analysera teman i en text finns om man kodar meningars början och slut, på så sätt att interpunktionstecken har särskild kod. Det är

antagligen det enda tänkbara sättet vid det slaget av inmatning. Vi är också av den uppfattningen att temat är väsentligt; det är nämligen där hela meningen får sin betydelse (att räkna substantiv som ett mått på ett visst innehåll kan ju knappast vara tillförlitlig information, om satsen eller meningen uttrycker förmodande, negering eller liknande). Vi kodar också temat, vilket de Inquirer-baserade systemen inte gör. Vad vi har vunnit är dels att vi kan bygga upp lexikon med begrepp som bas (på samma sätt som Pisa-konventionen förordar när man talar om ordet som bas), samtidigt som vi har varje satstema redan i inmatningen. Dels kan vi arbeta med delar av texten, t ex de olika teman för sig och behöver inte det tidsödande arbetet att söka ett tema i varje mening genom hela texten, på basis av vissa ord, t ex 'inte', som uttryck för negation.

Ruta 1. Exempel på autentisk text i intervjumaterialet och preparering för kodning

Frågan handlar om källor för kompletterande kunskap till det undersökta problemområdet.

Svar: ... Och, ja, sen hade vi TV-inspelningarna, där blev vi också hjälpta ganska mycket. Och så läsning. Det är väl dom huvudsakligen, dom källorna. Men det var väldigt bra att vara ute i fältet, faktiskt, det var det. - Nej, vänta lite grand, sedan så var det någon, det var innan min tid, va, det var före min tid, det var på hösten - 68, då så hade A, tror jag, och B gått ut med en enkät till forskare, som - nej, fastän den användes nog aldrig, nej - dom tänkte göra det i alla fall, gå ut med en enkät till forskare, där forskare på institutioner och handledare lite runt om i Sverige skulle tala om vad dom menar med självständighet, men jag tror inte den gav så mycket, för vi använde oss bestämt aldrig av den. Men vi hade kunnat göra något liknande --- för sent ute bara.

..//Och, ja, sen hade vi TV-inspelningarna,//där blev vi också hjälpta ganska mycket.//Och så läsning.//~~Det är väl dom huvudsakligen, dom källorna.~~//Men det var väldigt bra att vara ute i fältet, faktiskt, det var det.//- ~~Nej, vänta lite grand, sedan så var det någon, det var innan min tid, va,~~//det var före min tid, det var på hösten -68, ~~då så hade A, tror jag, och B gått ut med en enkät till forskare, som - nej, fastän den användes nog aldrig, nej~~ - dom tänkte göra det ~~i alla fall,~~ gå ut med en enkät till forskare, där forskare på institutioner och handledare lite runt om i Sverige skulle tala om vad dom menar med självständighet// men jag tror inte den gav så mycket// för vi använde oss bestämt aldrig av den.// //Men vi hade kunnat göra något liknande//--- ~~för sent ute bara.~~

Teckenförklaringar

Strykning av text ⊢—⊣
Supplering skall ske ———
Meningsavgränsning //
Interpunktion kodas ej

En begränsning i GI-systemet till följd av att hel text matas in utan föregående preparering, är att en hel del av det som har att göra med textförståelse av pragmatiskt slag går förlorat. Pronomen är ett exempel. Vad eller vem ett pronomen refererar till framgår inte när datorn sorterar orden. Vi har infört en suppleringsteknik, där referensen anges direkt efter begreppet i parentes (se I. Bierschenk, 1974). Detta illustreras i samband med presentationen av begreppskodning.

Antingen meningen eller ordet tas som bas finns problemet med idiom, när man matar in hel text. Vad som är idiom och inte är svårt att specificera med hjälp av t ex idiom-ordlistor. Arbetet med GI har bl a inneburit att skapa rutiner för hur datorn känner igen idiomen ("key-word-in context-procedure"), men detta förutsätter att det finns ett begränsat eller känt antal idiom, vars struktur likaså kan förutses. Idiom-ordlistor måste betraktas som otillräckliga, eftersom såväl talare som skrivare lätt bildar egna idiom, som en sådan ordlista omöjligt kan täcka. Problem med sådana pragmatiska drag i naturligt språk kan undvikas om kodning sker på basis av begrepp och inte ord. Till begrepp kan knytas flera ord, s k ordsträngar, som bildar egen "tag" (ett enskilt item) och datorn känner då igen en sträng med en specifik följd av ord.

Användaren av GI-systemet får ingen syntaktisk information, vilket innebär att t ex subjekt-objektrelationer eller rollfunktioner inte framgår av en utskriven lista av ord. Om således en analys av ett politiskt dokument skall göras där meningar som 'USA anföll Vietnam' och 'Vietnam anföll USA' ingår i texten, blir de betraktade som "innehållsligt" lika genom den procedur som endast identifierar ord och räknar frekvensen. Vårt analyssystem har skapat koder för syntaktisk funktion, som tar hand om sådana skilda satser som de exemplifierade och specificerar vem som anföll vem i de två fallen. För detta kodsystem har ingående redogjorts i I. Bierschenk (1974), men presenteras åter i väsentliga delar i ett senare avsnitt i denna rapport.

Redan 1966 ansåg Stone och andra att syntax-kodning är nödvändig för en förfinad analys "in the future". Men inte ens 1974 års förbättrade Inquirer har presenterat ett sådant system.

Antingen man kallar sin analysenhet mening, begrepp eller ord förordar Pisa-konventionen att ett item skall vara identifierbart. Denna förutsättning för att kunna återföra begrepp till ursprungstexten tillgodoses i vårt fall genom att varje begrepp har separat kort och därmed eget "nummer".

3. 2 Inmatning

Vår kodning sker på datablankett och har utformats så att vissa kolumner är specificerade för vissa typer av koder. Numerisk information i början eller i slutet av kortet, som konferensen rekommenderat, har tillgodosetts så att viss numerisk stansning sker i början och viss annan efter texten.

3. 3 Identifikation

Vår analys består av intervjusvar, där intervjuarens tal inte skall analyseras. Det är endast intervjupersonens åsikter som återspeglas i textföljden. Men det kan hända att den intervjuade refererar till vad någon annan sagt eller tyckt och sådan information har vi velat särskilja och har skapat en särskild kolumn som vi kallar "källa" med två alternativa koder, 1 för intervjupersonens egna åsikter och 2 för någon annans.

Figur 1 visar en kodad mening, där numeriska tecken i form av identifikation, tema och syntaktiska (eller funktions-) koder framgår. Uppskrivningen på blanketten görs så tidsbesparande som möjligt, så att samma kod nedåt i kolumnerna inte upprepas. Vid stansningen upprepas koderna såvida inte ny kod avbryter. Hur varje kort ser ut framgår av figur 2. Figuren visar också tydligt hur varje kort (begrepp) går att hänföra till texten genom att koderna upprepats vid stansningen. Angående de syntaktiska koderna skall här endast sägas att olika satser markeras genom looping till andra kolumner. Loop-koderna (fig 1) markerar i vilken satskolumn meningen fortsätter.

3. 4 Rekommendationer från Pisa

De rekommendationer (se s 13) som konferensen ansåg viktigast gäller uppbyggnad av register eller lexikon. En innehållsanalys av det här slaget bygger alltid på ett lexikon, men principerna för hur det byggs upp kan variera.

Den första rekommendationen avser särskiljande av typ (av ord) och fall (av samma ord). Svårigheter rapporterades med GI ifråga om distinktionen mellan typ och fall. Vi har i hög grad undvikit detta problem genom att i förväg koda syntaktisk funktion. Ett begrepp som t ex 'bibliotek' (vanligt i vårt provmaterial) kan vara subjekt (agent), objekt eller rumsbestämning. Den deskriptor som 'bibliotek' får kan alltså variera, vilket betyder att varje typ innebär ny syntaktisk-funktionell roll och därmed nytt innehåll. Vi har sedan möjlighet att räkna frekvens per typ, om vi finner det nödvändigt.

Den andra rekommendationen gäller bl a suffixbildning. Här ligger

PROGRAMMERINGS- OCH DATABLANKETT

Programmerare			Teckenstandard	EBCDIC / BCD	Speciell stansinstruktion
Kundnummer	Telefon nr	Datum			Sida av / Program

Assembler language: Namn 8 | 10 Operation 14 | 16 Operand | Kommentar
FORTRAN: Fml nr 5 | 6 | 7 FORTRANformel
DATA ALGOL PL/I Sats Sats
Identifikation / Sekvens

Kol. 1–10	11–15	26–65	66–72
3808003201	1 11	Man(jag)	30
2		vet	40
3		väldigt lite	45
4		om	40
5		vad(forskning)	50
6		som(forskning)	53230
7		händer	40
8		i det här landet(Sverige)	44

Identifikation
1–2 intervjupers. nr
3–4 fråga nr
5–8 mening nr
9–10 begrepp nr i men.
11 källa

Tema
12–14 negation, tempus resp modus (obl. kodning)
15–23 olika teman kodas vid förekomst

Begrepp
26–65 textlängd
24–25 * används när text löper över flera kort än ett

Syntaktisk funktion
66–80 funktionskoder
68,71,74,77,80 loop

Figur 1. Datablankett: Disposition och kodning

M7102-620070. Utgåva 1: 1970 april. 1000 bl à 100 ex (FOA 404)

```
38080032011 11      MAN(JAG)                      30
38080032021 11      VET                           40
38080032031 11      VÄLDIGT LITE                  45
38080032041 11      OM                            40
38080032051 11      VAD(FORSKNING)                50
38080032061 11      SÖK(FORSKNING)                50290
38080032071 11      HÄNDER                        40
38080032081 11      I DET HÄR LANDET(SVERIGE)     44
```

Figur 2. Hålkort: Stansad text

svenska språket närmare tyskan än engelskan och ordbildningen (sammansättningar) sker på samma sätt. Vi har under en tid arbetat med registerkonstruktion, vilket närmare redogörs för i kapitel 4. I det här sammanhanget skall endast nämnas att vi inte betraktar sammansättningar som "två ord", eftersom vi inte baserar vår analys på ord. En sammansättning som "socialforskning" blir hos oss betraktat som ett begrepp med annan betydelse än vad de två separata begreppen skulle ha, såtillvida att "forskning" är generellt och kan innefatta "socialforskning", där "social" är en bestämning. Skulle vi i en sökningsprocedur vara intresserade av begreppet "social" i de varianter det kan förekomma, specificerar vi denna sökning genom ett trunkeringsförfarande (se vidare kap 4.2).

Vi kommer inte att få problem med långa register med fallen registrerade i form av ändelser. Ändelserna specificeras för varje kategori av begrepp (vi håller t ex isär substantiv och verb) och finns i ett särskilt register. Häri liknar vårt system närmast GI s k "hash-list"-sökning (se sid 14).

I följande kapitel redogörs för den kanske viktigaste aspekten vid registeruppbyggnad, nämligen förfaringssättet med ändelser. Redogörelsen ges ett eget kapitel, dels för att belysa vilka möjligheter vi har att behandla våra begrepp, dels och främst för att visa att man även på svenska kan utarbeta register för vidare datorbaserade analyser. I vilken mån oklarheter ännu finns kommer att framgå.

4. UPPBYGGNAD AV LEXIKONBAS

4.1 Urval av begrepp

Begreppen är i kodsystemet indelade enligt funktion (och ofta position) i satsen. I likhet med traditionell satsanalys betyder det att det i subjektsposition kan stå ett verb, men att ett begrepp som till ordklassen är substantiv kan ha verbfunktion, även om sådana ordklassbyten inte är den vanligaste representationen. Nedan skall redogöras för på vilket sätt vi kan ta ut våra begrepp ur det inmatade materialet.

Om vi vill veta hur många och vilka adjektiven är beställer vi ut ur materialet de koder som innebär attribut med innebörden beskrivning eller klassificering av ett huvudord. Alla handlingar registreras i verbkod. Substantiv kan finnas i agent- eller objekt-koderna, men även i sådana koder som ligger utanför paradigmet A a O, t ex inom rumsadverbial. Nu har vi i första hand varit intresserade av att studera koderna inom paradigmet, så specificeringen av substantiv blir begränsad till agent, objekt eller mål.

Ett begrepp kan bestå, som tidigare framhållits, av flera ord, vilket ofta betyder en ordsträng, där huvudordet omges av artiklar och prepositioner. Detta gäller framför allt substantiven. För att få en bas för register av de begrepp som förekommer inom de olika koderna behövde vi först låta skriva ut de koder vi ville ha, sortera dem när partiklarna borttagits samt genom trunkering på olika sätt skapa möjligheter för sökning på dessa begrepp i dels olika böjningar och sammansättningar, dels olika koder, där de kan tänkas förekomma. Hur begrepp kan begränsas genom trunkering redogörs för nedan.

4.2 Trunkering: Syfte och principer

Begreppet trunkering kan sägas innebära att en avgränsning av element inom ett ord eller en viss specifik koppling av ord inom ett begrepp sker. När man vill söka sig fram till relevant information inom stora datamängder, t ex litteratur inom ett referatorgan, är man intresserad av att de termer som man söker genom är så entydiga som möjligt, så att inget brus uppstår i form av överflödsinformation. B. Bierschenk (1973, s 69) förklarar begreppet trunkering på detta sätt:

"För att skapa flexibilitet har man infört ett förfaringssätt som kallas trunkering, vilket innebär att man vid en maskinell sökning kan utnyttja ordstammar.

Genom att man kan trunkera ett ord på såväl höger som vänster sida blir även inom ramen för en automatiserad sökning en utskrift av titlar trots varierande prefix och suffix möjligt."

För att bygga upp begreppens register kan man sålunda utnyttja likheter och olikheter inom och mellan begrepp. Dels kan man undvika att behöva arbeta med olika böjningsändelser, som utan ett trunkeringsförfarande skulle betraktas som "olika ord" av en dator, dels kan man genom att enbart arbeta med ordstammar (främst i sammansättningar) "gardera" för att samtliga i ett material förekommande sammansättningar återvinns i en sökning. På vilket sätt detta sker illustreras genom trunkering av begreppet 'beteende' tillsammans med de kombinationer som SÖK-projektets material hittills skulle utvinna. (Trunkeringen utmärks med asterisk.)

Ruta 2. Resultat vid trunkering av begreppet beteende

Sökord	beteende*
	beteende*n
	beteende*t
	beteende*sidan
	beteende*vetare
	beteende*vetenskap
	beteende*modifikation
Sökord	*beteende
självständighets	*beteende

Högertrunkering betyder alltså att ordet till och med asterisken samt allt efter är relevant, vänstertrunkeringen garderar för ett föregående sammansättningsled. En dubbel trunkering (*beteende*) innebär därmed att samtliga begrepp utfaller i en och samma sökning, som ruta 2 visar.

I utfallet efter högertrunkeringen visar det sig att såväl ändelser (-n, -t) som begrepp innefattas i detta. Om man önskar begreppet 'beteende' utan efterled bör en rutin införas som först tar hand om de tänkbara ändelserna hos varje begrepp. Vi måste således ha ett ändelseregister innan vi genom trunkeringsförfarande kan söka igenom materialet på begrepp och begreppskombinationer. I detta första stadium har endast högertrunkeringar gjorts, vilka enbart avsåg ändelser, varvid samtliga möjliga ändelser registrerades. (I annat fall blir registret begränsat till det material vi hittills har bearbetat.)

Vi har utarbetat register för adjektiv, verb och substantiv såsom framgår av följande beskrivning.

4.2.1 Trunkering av adjektiv och verb

Generellt har här utgåtts ifrån att ett begrepp skall trunkeras så att så få ändelser som möjligt skall behöva registreras. Som ändelser betraktas därför här det eller de element (vi kan kalla det bokstäver) som om de förekommer efter asterisken ger en annan betydelse (i genus, kompara-

tion eller tempus) åt begreppet. Enkelt uttryckt kan vi säga att så långt som begreppet utan dessa element stavas lika, så långt trunkeras det. Men där stammen i begreppet har omljud, dvs annan vokal, registreras begreppet lika många gånger som antal stamvokaler. Ruta 3 visar principerna för trunkering av adjektiv och verb.

Ruta 3. Exempel på trunkering av adjektiv och verb

Adjektiv	Utfall/ändelser
stor*	-t, -a
stör*	-re, -st, -ste/a
fin*	-are, -ast, -aste, -a, -t
hjälpsam*	-mare, -mast, -maste, -ma/e, -t
Verb	
arbeta*	-r, -de, -t, -d[1], -s, -ts, -des
bjud*	-a, -it, -en[1], -s, -es, -its
bjöd*	-s
slå*	-r, -s
slog*	-s
slag*	-it, -its, -et[1], - en

1) participform (ibland som adjektiv)
s-formerna innebär att vi garderar även för passiver (möjligheten att säga "det bjöds" finns)

I en nästa uppbyggnadsfas kommer tempusformerna att sammanföras till ett begrepp som representerar de övriga. Vi har temakoder, som specificerar tempus i varje sats. (Jfr systemet WORDS, s 14.)

4.2.2 Trunkering av substantiv

Ändelserna för substantiv blir naturligt nog fler än för adjektiv och verb på grund av de många deklinationernas olika böjningsmönster och bestämd och obestämd form i olika genus. Dessutom har vi även här s-former som avser genitiv och som kan förekomma i samtliga andra former.

Även här förekommer pluralböjning med annan stamvokal, vilket behandlas på samma sätt som vid adjektiv och verb.

Ett av svenskans vanligaste avledningssuffix till substantiv är -ning. Detta betraktas inte som ändelse här, eftersom vi i så fall skulle få flera betydelser på en trunkering, som t ex skulle gripa över såväl abstrakta som konkreta begrepp, vilket skulle göra det svårare vid uppbyggandet av registren. Som exempel kan anges 'bokförsäljare,' 'bokförsäljning;' en trunkering 'bokförsälj*' är alltså överlappande inte bara med hänsyn till böjningsändelser. Ändelser betraktas på samma sätt här som ifråga om adjektiv och verb. Principerna vid trunkeringen behöver inte ytterligare presenteras. Exempel på vilken betydelse avledningsändelser har jämfört med böjningsändelser skall dock ges. Ruta 4 beskriver således substantivtrunkering som är garderad för avledningsskiljaktigheter.

Ruta 4. Exempel på trunkering av substantiv: Avledning jämfört med böjningsändelser

Trunkering	Utfall/ändelser
arkivarbet*	1) -e, -et, -ets, -en, -ena, -enas 2) -are, -ares, -arens, -arna, -arnas
arkivarbete*	-t, -ts, -s, -n, -na, -nas
arkivarbetar*	-e, -es, -ens, -na, -nas

Det torde vara självklart att endast den senare typen är tillfredsställande. Även liknande fall skulle kunna diskuteras, där skillnaden genom avledningsändelsen tycks mycket liten, t ex anmälan och anmälning. En viss skillnad finns dock och här har det genomgående utgåtts ifrån att det är fråga om begrepp med olika betydelse.

Här har diskuterats trunkeringsprinciper på basis av begreppet ordklass, trots att våra begreppskoder inte i första hand är ordklasser. Men ändelseregistren har vi velat bygga upp så att vi kan specificera ett sökord till att gälla ett begrepp inom en ordklass och inom en kod. I det fall ett verb står i agentposition (t ex "subjekt" i form av att-fras) och vi använder den trunkerade formen för ett verb vi vill ha ut, kan vi genom att specificera den syntaktiska koden undvika vissa oönskade utfall. Vi har alltså dubbel gardering. Trunkeringsproblem kan illustreras med två begrepp, 'ansvara' respektive 'ha ansvar för'. Dessa kan t ex båda vara verb, som i vårt system har kod 40 (se fig 1). Om vi i verbregistret har begreppet 'ansvara*' och gör en sökning med detta och samtidigt specificerar det till kod 40, får vi inte ut 'ha ansvar för' som verb, eftersom -a finns med till vänster om asterisken. Hade vi däremot haft 'ansvar*' hade båda begreppen utfallit, eftersom -a är verbändelse (se ruta 3). Den senare trunkeringen förutsätter att de båda varianterna för "samma" verb har samma betydelse. Om så skulle vara fallet och den sista trunkeringen gälla, betyder det att principen för trunkering av enkla verb kunde komma att ändras, så att -a faller utanför asterisken, men det syns för närvarande som en mindre bra lösning, framför allt med tanke på att kanske inte alla substantiv med verbbetydelse generellt har samma innehåll som sitt respektive verb. Sådana praktiska överväganden vid registerkonstruktion är principiellt viktiga. Vi är ännu inte helt färdiga med denna konstruktionsfas. Vi väntar på att få veta om de ovan diskuterade verben kan sammanfalla eller bör representeras i registret som var sin typ.

Distinktionen mellan typ och fall som diskuterades i Pisa gäller således inte enbart syntaktisk betydelseskillnad. Den eventuella skillnaden har vi specificerat genom kodningen. Möjlig skillnad i varianter är för oss väsentligare, vilket GI-systemet inte alls diskuterar. Den frågan är

inte aktuell när man baserar sina lexikon på ord. Ord blir med nödvändighet kategoriserade i ordklasser, oavsett funktion i kontexten.

4. 3 Empirisk bas

Hittills har varje begrepp betraktats som ensamt och olikt varje annat begrepp. Vilka som skall anses som lika, dvs ha samma betydelse och således genom ett trunkeringsförfarande endast en gång representeras i ett lexikon är ännu inte klarlagt. "Betydelse" blir i det här fallet bestämd genom skaleringstekniken med Osgoods semantiska differentialer. Samma teknik har använts i GI för att bestämma ett värde ("value") för de ord man vill använda i sin analys. Principen är i korthet den att adjektiv och verb får ett värde på en 7-gradig skattningsskala inom tre dimensioner (pos-neg, akt-pass, stark-svag) genom att ett antal bedömare skattar dem. Dessa medelvärden är det "value" som används för bestämning av substantivs innehåll, på så sätt att substantivs frekvens multipliceras med medelvärdet för det adjektiv eller verb som står tillsammans med substantivet i en text. På vilket sätt vi kommer att använda denna metod skall redovisas vid ett senare tillfälle. Skattningar har genomförts och statistiska beräkningar föreligger men är ännu ej analyserade. Dessa resultat är av vikt för registeruppbyggnaden i bl a de fall som diskuterats ovan. Kan vi komma fram till att två begrepp har lika betydelse förs deskriptorer enbart till detta begrepp med hänvisning till varianter som återfinns i ett särskilt register.

Genom att empiriskt kunna bygga upp registren anser vi att vi uppfyllt Pisas rekommendationer långt utöver vad som avsetts. Våra lexikon kommer att innehålla distinktionen mellan typ och fall, dels empiriskt förankrad ifråga om adjektiv och verb, som ju skall bilda basen för en kvantifiering av substantiv, dels en syntaktisk-funktionell distinktion, som vi kan använda oss av för utvidgningar och begränsningar i sökningarna, beroende på vilken problemställning vi har, t ex "Vilken värdering uttrycks i de handlingar som utförs av X"? eller "Hur värderas den handling som innefattar X som objekt"?

Vi kommer att få lexikon som är uppbyggda enligt förhållanden som gäller svenska språket och dessa lexikon utarbetas med det autentiska materialet som grund. Problemet med sammansättningar behöver vi inte ta ställning till, eftersom vi låter hela begrepp ingå i lexikonet och vi ändå kan söka sammansättningar genom trunkeringar.

Det torde stå klart efter denna schematiska presentation av lexikonuppbyggnad att det går att utarbeta regler för registrering med svenska språket som bas och att det är nödvändigt för varje innehållsanalysforskare

att skapa sin egen lexikonbas beroende på vilken språklig teori han har. Pisa-konferensen visade hur olika problemställningar ger nya problem med den språkliga representationen. Att bygga på det enskilda ordet är egentligen inget uttryck för en språkteori, snarare ett övervägande som gäller vad som är praxis vid datorbearbetning av text. Det enskilda ordet blir sen kategoriserat enligt någon samhällsvetenskaplig teori, där det enskilda ordet betraktas kontextfritt.

I vår analys är begreppen kontextbundna enligt en teori om relationer mellan beroende och oberoende begrepp och rollfunktioner. Det är efter dessa förutsättningar kodningen sker och vidare analys skall bli möjlig. En kodning av naturligt språk är en lyssnarföreteelse. Vad datorn skall åstadkomma är inte en förståelse av text utan en strukturering av texten såsom vi specificerat input.

Följande kapitel skall behandla den teori för språklig representation som vi har funnit vara den mest tillämpliga i vårt fall. Det diskuteras också sådana fall, där vi har låtit praktiska överväganden gå före teorin.

5. REPRESENTATION OCH KODNING AV BEGREPPSSTRUKTURER

I föregående kapitel har presenterats våra analysenheter, på vilket sätt inmatning sker, hur identifikation specificeras i kodningen och på vilket sätt vi har tänkt oss att bygga upp vår lexikonbas. I detta kapitel skall presenteras hur våra begrepp representeras genom kodningsregler (kolumnerna 66 - 80, fig 1).

Den teori som vi funnit mest intressant för vår analys är Schank's (1973) "conceptual dependency theory" som närmast kan hänföras till området matematisk lingvistik. Denna teori kom vi i kontakt med efter att våra första kodningsregler utarbetats, men vi fann att den ligger väl i linje med vårt sätt att behandla texten för input och den kommer att vara vägledande i fortsättningen, dock med vissa inskränkningar av praktisk natur.

5.1 Teoretisk ram

En analys av språk måste röra sig på två nivåer, dels en meningsnivå ("sentential level"), dels en begreppsnivå ("conceptual level"). Den första bildar basen med meningen som största enhet, den andra innebär begreppskomplex ("conceptualization") och satser med begrepp som enheter (se Schank, 1973, s 191). Det finns tre elementära slag av begrepp, "a nominal, an action, or a modifier". Dessa begrepp är antingen oberoende eller beroende. Nominal, dvs substantiv, är oberoende, alltså självständiga begrepp som inte behöver något tillägg för att förstås. Alla andra är beroende, dvs de måste relateras till andra begrepp för att få fullständig betydelse. Med modifier menas antingen adverb eller adjektiv. Verbet ("action") betraktas som "oberoende", men dess betydelse specificeras genom det eller de substantiv det sätts i samband med. Nominalen kallas hos Schank PP (för "picture producer"), verben ACT, adjektiven PA ("picture aiders") och adverben AA ("action aiders"). Dessa förkortningar används i fortsättningen vid jämförelsen mellan denna teori och vår analys. Egenskaperna hos begreppen gäller den conceptuella nivån och inte meningsnivån. Man kan ju ha en mening utan verb eller ett adjektiv utan sitt substantiv och yttrandet är ändå fullt förståeligt och betraktas som en mening i kommunikativt sammanhang. Det gäller däremot inte på begreppsnivån. Där gäller beroenderegler mellan begrepp, som indikerar att ett beroende förutsätter att ett oberoende existerar.

Detta förhållande mellan mening ("sentence") och mening ("meaning") är den väsentligaste frågan när man vill utveckla ett datorbaserat system för innehållsanalys. Vi vill analysera innehåll via begrepp och deras relatio-

ner och inte syntaxen i sig själv. Schank (1973, ss 189-190) framhåller följande:

"Regardless of the purpose, however, it is the meaning of an input sentence that is needed not its syntactic structure. /. . . / rather than doing a complete syntactic analysis and then submitting the results of this analysis to some semantic interpreter, we would like to use syntactic information as a pointer to the conceptual information."

5.2 Representation och kodning av begrepp

En "conceptualization" uttrycker en händelse och kräver således ett verb och minst två substantiv. Hur en sats tolkas beror på de conceptuella reglerna. Det existerar formellt definierade beroenderelationer mellan givna kategorier av begrepp. Det är dessa beroenden som utgör strukturen på begreppsnivån.

Schank har utvecklat ett s k C-diagram ("conceptual dependency network") för att symboliskt uttrycka beroenden mellan begrepp. Sex grundregler finns för ett sådant nätverk och dessa beskrivs nedan (ur Schank, 1973, s 195).

Ruta 5. Exempel på conceptuella regler

C-regel	Exempel
1. PP ⟺ ACT	John ⟺ hit (←o— Mary)
2. PP ⟺ PA	John ⟺ tall
3. PP ⟺ PP	John ⟺ doctor
4. PP ↑ PA	John ↑ tall
5. PP ⇑ PP	man ⇑ N.Y. LOC dog ⇑ John POSS-BY
6. ACT ←o— PP	(John ⟺) hit ←o— Mary

Regel 1. Denna regel säger att PP kan handla och när de gör det existerar ett ömsesidigt beroende mellan dem. Regeln säger däremot inte att alla PP kan handla eller att vilken PP som helst kan utföra alla handlingar. Regeln rör alltså syntaxen på begreppsnivån. (Semantiska regler skulle tala om vilka PP som kunde utföra vilka typer av ACT.)

Regel 2. Det är möjligt att predicera ett attribut till ett speciellt PP. Denna relation indikeras med ⟺ . För att en "conceptualization" skall existera mellan dessa två items är båda lika nödvändiga och beroendet är alltså ömsesidigt.

Regel 3. Man kan också predicera "gruppgemenskap" ("set membership") mellan två PP.

Regel 4. Man kan referera till ett begrepp och ett attribut till detta begrepp, vilket redan tidigare nämnts i kontexten.

Regel 5. Två conceptuella objekt kan relateras till varandra på olika sätt. Principiellt finns tre sätt: innehåll ("containment"), plats ("location") och tillhörighet ("possession"), markerade genom ⇑ . Begreppet under pilen specificerar (är beroende till) begreppet över.

Regel 6. Regeln anger objekt-beroende. ACT fordrar ett PP. En given ACT kan kräva ett objekt, andra gör det inte. De som gör det måste ha ett objekt på begreppsnivån, även om det inte uttrycks på meningsnivån.

När det gäller att på ett systematiskt sätt använda dessa principiella regler för inmatning av text i dator måste man finna på ett sätt att genom kodsystem representera begrepp och begreppsrelationer. Väsentligt är att påpeka att semantiska regler eller tolkningar inte i första hand skall uttryckas genom kodningen på den här nivån. En tolkning av texten är dock ofrånkomlig; i annat fall kan man ju inte heller segmentera begreppen. Att därför påstå att syntax och semantik skall särskiljas är inte hållbart, eftersom syntax och semantik är varandras förutsättningar. Vi har använt oss av syntaxen för att med ledning av denna struktur kunna avgränsa begrepp i en sats eller mening.

Som tidigare nämnts vill vi huvudsakligen bygga upp våra meningar enligt paradigmet A a O. Vi skall här försöka visa hur förhållandena mellan dessa beteckningar symboliskt kan kodas. Dessutom har vi genom kodsystemet försökt ange beroendestrukturen inom begreppskomplexen. På vilket sätt vi ansluter oss till Schanks språkteori redovisas i ruta 6.

Ruta 6. Jämförelse mellan C-diagram och ANACONDA: Begreppskodning

C-diagram	ANACONDA
1. PP ⟺ ACT	30 + 40
2. PP ⟺ PA	30 + 41 + 32
3. PP ⟺ PP	30 + 41 + 30
4. PP ↑ PA	32 + 30
5. PP ⇑ PP	30 + 33 31 + 30
6. ACT ⟵o PP	40 + 50

Vårt kodsystem är uppbyggt så att varje begrepp specificeras genom en tvåsiffrig kod. A a O uttrycks som 30 + 40 + 50. Kod 30 innebär agent- eller subjektfunktion. (Vi anger inte om agent är person eller abstrakt begrepp. En senare kategorisering av respektive kod kommer att ta hand om sådana bestämningar.) Ett begreppskomplex består av huvudord och bestämningar av olika slag. Huvudbegreppet har en kodsiffra som slutar på 0, medan beroende begrepps slutsiffra är 1, 2 eller 3.

Förhållandet mellan PP och PA uttrycker beroende, där PA är beroende begrepp till PP. Detta slag av beroende inom PP-strukturen uttrycker vi genom sifferkombinationen (30 + 32). Regel 2 säger att den "conceptualization" som bildas fordrar ett ´är´. Kopula-verb särskiljs hos oss från handlingsverb och får kod 41. Det gör det också lätt för oss att i materialet urskilja alla meningar som uttrycker värderingar och klassificeringar i form av predikativa påståenden.

Regel 3 fungerar på samma sätt, men här är utfyllnaden inte ett attribut (adjektiv) utan ett substantiv, lika oberoende som det första.

I stället för ´är´-konstruktion kan förhållanden mellan PPoch PA uttryckas såsom regel 4 visar. Adjektivet står framför sitt substantiv (t ex ´den långe mannen´). Schanks pil under PP betyder att detta begreppskomplex inte bildar en syntaktiskt fullständig mening, eftersom ACT saknas, dvs ett begrepp som står horisontellt i förhållande till PP. Detsamma gäller den femte regeln, som visar beroendet mellan substantiv inom ett begreppskomplex. Enligt Schank finns det tre slag av beroenden som uttrycks med denna regel, "containment", "location" och "possession". Vi kallar detta för framför- eller efterställd bestämning, som inte är adjektivattribut. Som 30 + 33 skulle vi koda alternativen ´mannen i New York´, ´ärterna i burken´; ´burkens ärter´ däremot som 31 + 30. ´Johns hund´ (se ruta 5) är ett ägandeförhållande uttryckt i genitivform, som i vårt fall kodas genom 31 + 30.

Regel 6 anger beroende mellan verbet och objektet. Schanks symbol säger att ett beroende finns, såtillvida att ett verb kan fordra ett objekt i en fullständig conceptualization, men att objektet inte för övrigt är ett beroende begrepp. Det är endast i den horisontella linjen som det i vissa fall är nödvändigt. Detta förhållande avser frågan om transitiva och intransitiva verb. Vi kodar objektet med en oberoende kod, eftersom ett PP i denna ställning som begrepp har samma struktur som ett PP som agent. I en mening kan nämligen finnas begrepp som hänförs till objektet i form av bestämningar av olika slag (se ruta 7). Hade vi tilldelat objektet en kod

som fick det att tillhöra verbet (någon siffra på 40) hade vi inte fått symmetri mellan agent respektive objektkomplexen. I stället skiljer vi på slaget av objekt. Det objekt som kodas med 50 är det som hör närmast till verbet, motsvarande det som traditionellt brukar kallas direkt objekt. (Ett andra objekt finns och detta presenteras längre fram.)

Innan framställningen går vidare med nya jämförelser exemplifieras vår kodning av en basmening ('John hit his little dog'), som Schank inledningsvis har presenterat, för att visa hur hans språkteori kan praktiskt användas i input.

Ruta 7. Jämförelse mellan C-diagram och ANACONDA: En basmening

C-diagram	ANACONDA	
John ⟺ hit ←c— dog ↑ (little) ⇑ POSS-BY John	John hit(s)[1] his (John's) little dog	30 40 51 52 50

[1]tempus är ej markerat i C-diagrammet

Ur rutan framgår på vilket sätt objektet behandlas. Det har samma grundstruktur som (agenten) subjektet om man utgår från deras sammansättning av oberoende och beroende enheter. För att understryka och hålla isär PP syntaktiskt har vi således det tvåsiffriga kodsystemet. Därigenom kan vi särbehandla begreppen på ett flexibelt sätt. T ex var det mycket lätt att ta ut alla adjektivattribut inför den skalering vi genomfört, eftersom de är specificerade genom den andra siffran i koden.

Förhållandet mellan de två PP 'John' och 'dog' måste få sin representation på ett otvetydigt sätt. Eftersom vi inte arbetar med automatisk igenkänning av items utan förbereder den maskinella databehandlingen manuellt, specificeras alla pronomen genom en supplering inom parentes. (Se vidare under kap 5.3.) Schank (1973, s 196) säger följande om förståelsen av naturligt språk:

"... in attempting to uncover the actual conceptualization underlying a sentence, we must recognize that a sentence is often more than its component parts. /.../ In fact, a dialogue is usually based on the information that is left out of a sentence but is predicted by the conceptual rules."

Den information som är utelämnad men som kan prediceras genom beroendestrukturen mellan begrepp är i detta sammanhang viktig för vår analysmetod. Pronomens referens är det första vi måste ta hänsyn till för att kunna arbeta med de begreppen. Utan referens till begreppen skulle ett mycket stort material gå förlorat, vilket också diskuterades i samband med GI. Emellertid är pronomens referens inte en huvudfråga i det här sammanhanget. Utfyllnaden av ett begreppskomplex (t ex inom agentstrukturen) eller de nödvändiga delarna i det syntaktiska paradigmet är teore-

tiskt intressantare i en teori om begreppsformationer eller image "underlying a sentence". Vi skall gå vidare och redogöra för vad Schank avser med underliggande strukturer.

Det centrala i Schanks resonemang är betydelsen i verbet. ACT innebär en händelse eller ett skeende som uttrycker rörelse eller stillastående. Ett rörelseverbs riktning brukar av språkforskare betecknas med transitivitet. En conceptualiseringinnebär i dessa fall att man vet målet för denna riktning eller att man känner till att ett mål bör finnas. Schank redogör för sin teori om TRANS genom ett antal meningar med verb som innefattar underliggande men nödvändiga kasus.

Ex. 'The man took a book' analyseras begreppsmässigt:



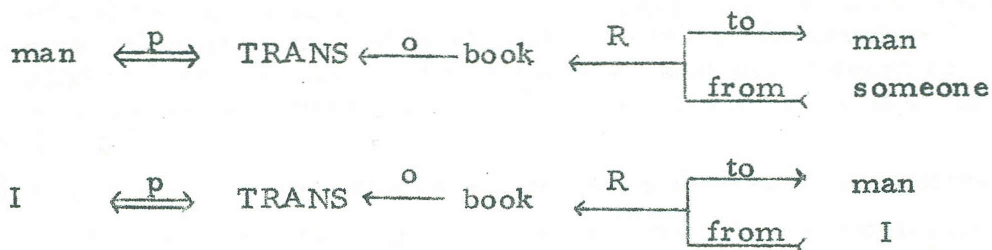
(p indikerar tempus, "past", o anger att verbet fordrar objekt.)

Denna mening är inte fullständigt representerad i och med att verbet take förutom objekt måste ha begrepp för "från vem eller vad" samt en mottagare. Nätverket ser ut så här:



(R står för mottagaren, "recipient". Detsamma gäller verbet 'give', men med x känt t ex 'from I'.)

De "conceptualizations" som ligger under meningarna 'The man took a book' respektive 'I gave the man a book' representeras på det här sättet:



'Give' definieras som TRANS när agent och källa ("originator") är identiska, 'take' är TRANS när agent och mottagare är identiska. Schank (s 198) förtydligar:

"This conceptual rule states that certain ACT's require a two-part recipient in a dependency similar to that of objective dependency. The similarity lies in the fact that this type of dependency is demanded by certain members of the category ACT. If it is present at all, it is because it was required. /.../ ... a conceptualization is not complete until all the conceptual cases required by the ACT have been explicated."

De fyra conceptuella kasus som Schank arbetar med är OBJECTIVE, RECIPIENT, DIRECTIVE och INSTRUMENTAL. Dessa kasus representeras i de C-diagram såsom ruta 8 visar. Regel 6 från ruta 6 upprepas här. Som

jämförelse presenteras på vilket sätt ANACONDA-systemet skulle koda dessa kasusförhållanden.

Ruta 8. Jämförelse mellan C-diagram och ANACONDA: Conceptuella kasus

C-diagram	ANACONDA
6. ACT $\overset{o}{\longleftarrow}$ PP	40 + 50
7. ACT $\overset{R}{\longleftarrow}$ [→ PP, ← PP]	30 + 40 + 50 + 70
8. ACT $\overset{I}{\longleftarrow}$ ⇕	40 + 45
9. ACT $\overset{D}{\longleftarrow}$ [→ PP, ← PP]	30 + 40 + 44 30 + 40 + 50 + 44

Regel 6 har redan förklarats.

Regel 7 säger att en handling som har ett PP som mottagare också måste ha ett PP som är initiativtagare eller agent förutom det PP som är kopplat till verbet, dvs en sådan handling fordrar två objekt. Vårt andra objekt har koden 70 och har alltså samma struktur som övriga oberoende begrepp i paradigmet. Vi har i kodningsreglerna kallat detta objekt för indirekt (enligt traditionellt sätt att analysera satser) eller mål ("goal").

Regel 8 visar att det kasus som kallas instrument genom den lodräta pilen kan betraktas som beroende till handlingen. Hos oss motsvaras det närmast av ett sättsadverbial.

Regel 9 förklaras av Schank (1973, s 202) så här:

"The DIRECTIVE case indicates that PP's may serve as direction indicators of a directional action. /.../ The directive case is extremely similar to the recipient in form and is almost in complementary distribution with it. That is, the two never appear together and would seem to be different forms of the same phenomenon. The most common ACT that take directive case is 'go'."

Kod 44 har hos oss specificerats som rumsadverbial, som uttryck för såväl befintlighet som riktning. Det visade sig dock vid kodningen att det var svårt att skilja mellan "riktning vart" och "mål". Detta är senare löst så att kod 44 används både som riktning och befintlighet ifråga om agentens läge. Om handlingen innefattar ett objekt, som byter "ägare" används koden för mål, dvs motsvarande recipient. Skillnaden förklaras genom följande två meningar:

Jag skickade papperen till institutionen
30 + 40 + 50 + 70

Jag reste till/ bor i/ Stockholm jfr Jag la papperen på bordet
30 + 40 / 40/ + 44 30+40 + 50 + 44

Hittills har vi inte skilt på riktning och befintlighet. Vill vi av någon anledning göra det behöver vi bara kopplingen mellan 40 och 44 för att se vilketdera som uttrycks i adverbialkoden. Observera att kod 44 inte uttrycker "location" enligt regel 5 (ruta 5), där LOC är bestämning. Exemplet med

30 + 40 + 50 + 44 betyder att agenten samtidigt med handlingen befinner sig i närheten av det PP som uttrycker plats.

Skillnaden mellan objekt och instrument kan vara besvärlig att reda ut. Schank (1973, ss 199-200) tar ett exempel:

'John grew the plants with fertilizer.'

Begreppet 'fertilizer' är det syntaktiska instrumentet till 'grew'. Men hur är det på den conceptuella nivån? Ja, han menar att verbet 'grow' inte kan vara en handling som en person kan utföra mot någonting. Det är växterna som blir större av det som John gjorde. Det är alltså fråga om en förändring av tillstånd, vilket måste uttryckas i en ny regel:

```
Regel 10.   PP ⇐┬──→ PA
                └──← PA
```

som, representerad i meningen ovan, blir:

```
plants ⇐┬──→ size = x + y
        └──← size = x
```

Johns handling representeras genom ett s k dummy-verb, 'do':

```
John ⇔ do ←─I── fertilizer
```

Därmed har vi fått två conceptualiseringar som måste relateras till varandra på något sätt, nämligen genom ett orsakssamband, som anges med ⇑ mellan de två satserna (i anger att orsakandet var avsiktligt, "intentional"):

```
John ⇔p do ←─I── fertilizer
      ⇑i
plants ⇐p┬──→ phys st size = x + y
         └──← phys st size = x
```

Egentligen är 'fertilizer' inte instrument utan objekt, eftersom följande antagligen hände: John tog gödselsäcken och gick fram till växterna och hällde ut innehållet på marken där växterna stod. Detta fick växterna att växa. På begreppsnivån är detta ett annat slag av TRANS:

```
John ⇔p TRANS ←─o── fertilizer ←─D─┬──→ plants ground
                                   └──← bag
      ⇑i
plants ⇐┬──→ phys st size = x + y
        └──← phys st size = x
```

Slutsatsen är att vad som såg ut att vara ett syntaktiskt instrument, dvs instrument på meningsplanet, är på begreppsplanet ett objekt. Detta händer alltid med ett syntaktiskt instrument, säger Schank, eftersom ett ensamt PP inte kan vara conceptuellt instrument utan bara objekt till en handling. Detta är förklaringen till regel 8 ovan.

På detta sätt fortsätter Schank att utreda underliggande strukturer i meningar. Representationen av meningen 'John ate ice cream with a

spoon' har ett ännu större nätverk på den conceptuella nivån i och med att verbet 'äta' innebär en serie handlingar som var och en har nödvändiga instrument, t ex 'ta skeden', 'röra handen med skeden mot glassen' osv. Schank menar att varje handling kräver ett instrumentellt kasus, men man behöver inte ange dessa vid verb som 'äta', där varje lyssnare vet vilka instrument som erfordras och vilka som kan vara möjliga. Man tänker inte aktivt på dem. Vid kodning av naturligt språk, på det sätt en lyssnare gör när han förstår vad en talare säger, finns alla handlingsserier implicita i ett sådant verb. Målet för oss är att ange vilket begrepp i satsen som indikerar handling. Underliggande handlingar och instrument är irrelevanta. Av den anledningen kan vi gå förbi regel 10, som säger att ett PP kan byta tillstånd. Allmänt är regeln riktig, men inte för oss i det här sammanhanget. En mening som 'Han gödde växter' kodas 30 + 40 + 50. Att växterna därav blir större får vi anta och den innebörden finns i verbet utan att det behöver anges med en relaterad sats 'så att växterna blev större' eller liknande. På samma sätt skulle vi göra i 'Han gladde mig'. Vari handlingen bestod vet vi inte och vi behöver inte heller veta det för att kunna representera begreppen i en kodning.

Den diskussion som förs angående skillnaden mellan objekt och instrument behöver inte bereda oss några problem. Vår kod 45 som anges i regel 8 kan ses både som instrument och objekt. Siffrorna i koden anger dock att begreppet hänför sig till verbet som modifierare, dvs inom 45 har vi sättsadverbial, som kan bestå av adverb eller substantiv. Innebörden i denna kod är: på vilket sätt utförs en handling, tillvägagångssätt eller metod? Koden har hittills också innefattat gradadverb och har fått hög frekvens, vilket motiverar att den delas. Det går dock inte att utgå från distinktionen substantiv-adverb, eftersom substantiv kan vara abstrakta begrepp som likaväl kan uttryckas genom ett adverb, som t ex 'med frenesi', 'frenetiskt'.

Skillnaden mellan Schanks instrumentella kasus och vårt adverbial kan sägas vara att instrumentet finns implicit i verbet, medan adverbialet kodas hos oss, om det finns explicit som modifierare till handlingen. Tyvärr diskuterar Schank inte adverbial (AA), vilket borde ha varit på sin plats här. Därmed är vi inne på frågan om de conceptuellt nödvändiga delarna på lämpligaste sätt kan kodas. Här skall därför redogöras för hur vi anser att begreppsnivån kan göras explicit för vårt syfte.

Det naturliga språket som informationsmedel kännetecknas av en ekonomi, som bl a innebär att referens uttrycks med pronomen. Ruta 7 visade hur vi kodar ett personligt pronomen med referenten inom parentes. Det finns även andra sätt att uttrycka en till innehållet fullständig mening

utan att alla nödvändiga delar är explicita. I ett samtal mellan två personer t ex behöver inte alla delar vara med, eftersom de i sammanhanget är kända för mottagaren av meddelandet. Principen kan visas med detta enkla fråga-svar-exempel: A: 'Vill du ha det stora eller det lilla äpplet?' B: 'Det lilla.' Vad B menar är: 'Jag vill ha det lilla äpplet.' Överfört på den tidigare diskuterade teorin om beroenden innebär kodningen av B:s svar att vi för det första kodar svaret som ett beroende begrepp (PA) som tillhör det oberoende (PP) 'äpplet'. Dessutom ligger i svaret den syntaktiska funktionen objekt ($\xleftarrow{o}$) och det räcker inte som conceptualisering. Verbet är 'vill ha' (vi måste utgå från det som tidigare sagts och inte vad B eventuellt hade sagt i stället). Men 'vill ha' + objekt (ACT $\xleftarrow{o}$ PP) är heller inte tillräckligt; agent (eller subjekt) 'jag' fattas.

Det som vi här har kallat vertikalt beroende gäller utfyllnaden av begreppskomplex, så att huvudord görs explicit i kodningen. Om huvudordet är det begrepp som sägs kan vi betrakta begreppskomplexet som självständigt, men i de fall attribut eller bestämningar står ensamma i satsen supplerar vi in det begrepp som är det självständiga och som återfinns i den föregående kontexten. Det kan också finnas fall där det begrepp som är huvudord inte tillräckligt förklarar hela komplexet, t ex 'lärarhögskolan'. Vi kan här utgå från ändelsen, den bestämda formen. Begreppet måste då ytterligare bestämmas, t ex 'i Malmö', som då är en efterställd bestämning med siffran 3 som andra siffra i koden. Vi har utarbetat regler för hur våra suppleringar skall göras (I. Bierschenk, 1974), men de är ännu inte fullständiga.

Vi skulle kunna kalla det beroende som uttrycks genom oberoende begrepps förhållande till varandra i en sats som horisontellt beroende. Liksom Schank utgår vi från verbet som det centrala begreppet i en sats. Verbets "riktning" avgör vilken kod de involverade huvudorden får. Det betyder att vi inte kan göra en syntaktisk kodning, där det första huvudordet (PP) är subjekt oavsett innebörden. Vårt paradigm A a O är vägledande och det PP som är agent behöver inte alltid vara satsens subjekt. Vi betecknar det substantiv som är styrande som agent, oavsett om det är abstrakt eller konkret. Det betyder samtidigt att vi inte i förväg anger om verbet är ett handlingsverb eller ett tillståndsverb. Däremot skiljer vi på kopula-verb och andra verb och därmed har vi gjort en distinktion mellan substantiv som utför handling eller befinner sig i ett tillstånd och de som är föremål för värdering eller klassificering från talarens sida. Bestämning av verbens aktivitetsgrad sker genom skaleringen och först därefter hoppas vi kunna bilda kategorier av verb. Med ledning av dessa resultat blir det också möjligt att gruppera substantiven enligt en innehållsteori på

empirisk grund. Vi innehållsbestämmer alltså inte våra koder mer än till den conceptuella funktionen och den liknar mycket Schanks kasusteori som bygger på syntaktisk roll.

Hur viktigt verbet är som nyckel till hur övriga begrepp skall kodas illustreras i ruta 9.

Ruta 9. Kodning av riktning vid passiv verbform

DET HÄR (X)	50
HAR	40
JU	42
ALDRIG	--
UNDERSÖKTS	40
FÖRUT	43

Kod -- betecknar negation

'Har undersökt' är den aktiva formen. Agent till verbet saknas och är här så pass odefinierad (ev 'av forskare') att den inte har supplerats in. Däremot är det klart att objektet är X och att detta inte kan vara agent. Verbets passiva form gör det omöjligt.

Agentlösa satser med passivt verb är tämligen vanliga, när talaren inte anser att agenten är väsentlig i sammanhanget. Den finns heller inte i kontexten. Vårt paradigm är inte uppfyllt här, men meningen kodas ändå. Vi har ansett det viktigt att så mycket text som möjligt läggs in och vi kan då i efterhand arbeta med de olika slagen av satser. Den information vi vill ha avgörs genom de kopplingar av begrepp som är tillgängliga. Verbet avgör således om paradigmet kan kodas i sin helhet eller inte. I detta fall är formen indicium. I andra fall beror kodningen på betydelsen i verbet. I ruta 10 visas två typer av verb, där det ena "fordrar" en supplering av det begrepp vi kallat mål, dvs som motsvarar Schanks kasus recipient, och det andra är en typ av verbbegrepp, vars huvudkonstituent egentligen är ett substantiv.

Ruta 10. Kodning av det fullständiga paradigmet

VI (XYZ-PROJEKTET)	30
SKICKADE IN	40
EN PROJEKTANSÖKAN	50
(TILL RIKSBANKSFONDEN)	70
JAG	30
VAR HANDLEDARE FÖR	40
LÄRARKANDIDATER	50
PÅ ÄMNESLÄRARLINJEN	53

I verbet 'skicka' ligger implicit ett mål eller en mottagare. När detta begrepp är känt i texten supplerar vi det så att en fullständig conceptualisering bildas. Det andra exemplet ges här som en illustration till att vi inte gör en

fullständig syntaktisk analys. En sådan skulle nämligen inte ha betraktat ´handledare´ som verb och förbindelsen mellan ´jag´ och ´lärarkandidater´ skulle inte ha framträtt. Vi ser ´var handledare för´ som ett verb motsvarande ´handleda´, till vilket det måste finnas ett objekt. Däremot finns här inget mål för handlingen handleda, som inte uttrycker riktning. (På vilket sätt vi definierar verb framgår av I. Bierschenk, 1974, s 57 och 1975, ss 8-9.)

5.3 Representation och kodning av relationer

I kapitel 5.2 har diskuterats förhållandet mellan beroende och oberoende begrepp och i vilket förhållande sådana begreppskomplex står till varandra i en fullständig conceptualisering. En fullständig conceptualisering kan emellertid i sin tur relateras till en annan, och detta förhållande motsvaras av satsfogning. Förhållandet mellan concepts kallas beroende ("dependency"), medan förhållandet mellan conceptualiseringar kallas relation.

Den viktigaste conceptuella relationen är enligt Schank (s 202) den som uttrycker kausalitet, vilket anges symboliskt:

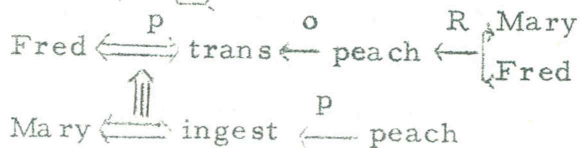


Kausaliteten uttrycks genom pilen mellan de båda satserna. (Skall ej uppfattas som logikens tecken för implikation.) Schank tar några exempel på hur kausalitet får sitt uttryck i engelskan:

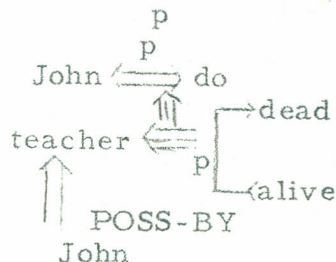
John was sad because Mary hit him.



When Fred gave Mary a peach she ate it.



John killed his teacher



(INGEST är en kategori av ACT liksom TRANS.)

I verbet ´döda´ ligger betydelsen av tillståndsförändring som leder till ett resultat enligt regel 10 (se sid 36). Handlingen i ´kill´ skulle kunna vara ´by shooting´, som då realiseras som t ex ´propel bullets via a gun to the

teacher's head'. 'Döda' är alltså en klass av transitiva verb som Schank kallar "pseudo-state-verbs". De har egenskapen att verbets objekt är 'actor' i den beroende conceptualiseringen (läraren dör). Ofta är verbet ACT i den beroende satsen.

'Sam flew his plane to San Francisco' skulle tydas 'Sam gjorde så att hans plan flög honom till S. F.' Det är alltså ett slags gömd kausalitet, som också diskuteras i exemplet 'John comforted Mary', där John orsakar ett tillstånd hos Mary (regel 10). Regel 10 har redan avfärdats som irrelevant för oss och det tredje exemplet ovan lämnas därhän i den fortsatta jämförelsen.

Regel 11 skall nedan jämföras med ANACONDAS kodningsregler.

Med utgångspunkt i ovanstående exempel på kausalitet, bör man ställa frågan: Vad menas här med kausalitet? Splittringen av regel 11 i a och b gör det möjligt att förklara detta som ett mer generellt fenomen än att relationen skulle gälla sådana satser som traditionellt skulle analyseras som orsakssatser. Kausalitet uttrycker relation mellan satser, där två (eller flera) handlingar åtskiljs i tiden. Schank (1973, s 205) säger:

"In other words, we want to be sure to distinguish distinct conceptual events in the real world."

Det första exemplet skulle analyseras som en mening med en orsakssats inledd med 'because' (på svenska 'eftersom', 'på grund av', m fl). Den andra meningen uttrycker snarare en förutsättning, där första satsen i tiden ligger före den andra. Sambandet ligger i att Mary åt persikan efter att ha fått den. Det tredje exemplet uttrycker en "dold" orsak, som beror på betydelsen i verbet.

I ANACONDA-systemet finns varje sats kodad för sig inom sin mening, på så sätt att satserna tar var sina två kolumner i anspråk. Vad som är sats avgörs alltså genom hur många verb meningen har, vilket är detsamma som en åtskillnad mellan "distinct conceptual events". Relationer mellan satser uttrycks inte bara genom en viss ordningsföljd eller ett visst antal, utan också genom det tema som hela meningen kodas enligt. Temakoder återfinns före texten, i anslutning till identifikationskoderna och betraktas alltså som överordnade begreppskoder. Ett tema gäller inte något enskilt begrepp, utan kodas genomgående till varje begrepp. Vi har då möjlighet att genom en kod i början av kortet sortera ett begrepp som tillhörigt ett specifikt tema. Var på blanketten temakoderna finns illustreras i figur 1. På vilket sätt Schanks två första exempel kan kodas presenteras i figur 3.

De koder som kan förekomma i temakolumnerna (15-23) är antingen 1 eller 2. 1 betyder att satsen kommer före i tiden i förhållande till den

Tema- koder		Text	Satskodning, "kolumner"	
21	2	John	30	
		blev	41	
		ledsen	32	
	1	eftersom	462	
		Mary		30
		slog		40
		honom (John)		50
21	1	När	462	
		Fred		30
		gav		40
		Mary		70
		ett äpple		501
	2	åt	40	
		hon (Mary)	30	
		det (äpplet)	50	

I kolumn 15 kodas villkor, 16 orsak, 17 medgivande, 18 följd/avsikt, 19 disjunktion, 20 komparation, 21 fråga, 22 antagande, 23 önskan

Figur 3. Kodning av relationer

andra satsen. Den syntaktiska följden bibehålls. Satskodningen visar hur satser löper i slingor. För detta system redogörs utförligt under figur 4. En siffra efter begreppskoden visar i vilken kolumn meningen fortsätter. Både 'eftersom' och 'när' anger ny sats och kallas för satsinledare, som oftast är ett adverbial; därav beroende kod till verbet. När stansning sker upprepas koden i temakolumnerna tills dess att den bryts av ny kod (gäller ej vid meningsslut). Kod 46 som satsinledare är indikator på någon form av relation som inte binder ihop enskilda begrepp utan hela satser. Schank (1973, s 206) uttrycker detta genom:

"A relation is used to relate dependencies not concepts."

Tolkningen av den första satsen är enkel, eftersom orsakssambandet uttrycks explicit, så det behöver inte närmare diskuteras. Den andra meningen kan däremot vara svårare. Den är inte lika entydig. Tolkningen i figuren innebär att meningen anger en handling som får en följd i en ny handling. Parentetiskt bör nämnas här att följd och avsikt har sammanslagits. Det är ganska svårt att i talspråkstext skilja på 'så att ...' och 'för att ...' - satser. Oftast uttrycks båda innebörderna med en så-konstruktion. I den här kodade meningen antas antingen att Fred hade för avsikt att Mary

skulle äta äpplet eller att Mary åt äpplet som en följd av att hon fick det (Han gav det så att hon skulle kunna ...). Det hade också funnits en annan möjlighet att koda den meningen, nämligen att se den första satsen som en tidsbisats. I stället för 46 hade då ´när´ fått 43, som uttrycker tidsadverbial, men temakod hade slopats. Det är i sådana fall som utbrytningen i satskolumner är till hjälp vid återvinningen. Huvudsatsen är den första, i det här fallet med ett fullständigt paradigm (30 + 40 + 50). Förutom paradigmet finns begrepp som uttrycker omständigheter av olika slag runt handlingen. Handlingen i denna bisats är ´gav´ och denna handling måste vara före i tiden, vilket är sådana bisatsers karaktär. Ett annat indicium är ordföljden i huvudsatsen. Till identifikationen hör ordningsföljden bland begreppen. Om kod 40 står före kod 30, kan vi prediciera att en bisats kommer före, om det inte finns temakod för fråga. Är däremot ordföljden i huvudsatsen den omvända, går det inte att göra en sådan prediktion, utan då får vi använda koden 43. Problemet med denna typ av relation behöver inte vara så stort. Antingen vi kodar satstema eller har explicit en tidsangivare framträder relationen; i bägge fallen gäller det ju handlingars tidsaspekt. Exemplet visar dock att det finns många omständigheter att ta ställning till vid utvecklingen av ett datorbaserat system för innehållsanalys. Paradigmet A a O skulle ensamt vara otillräckligt när det gäller kodning av satsrelationer.

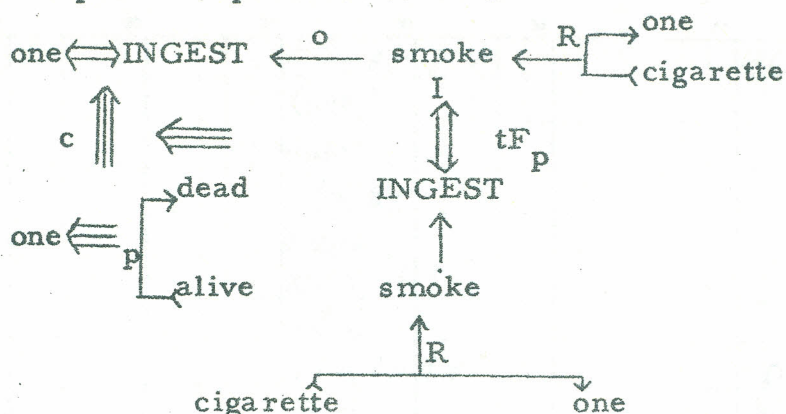
Vi har således koder för sådana aspekter i en mening som inte kommer till uttryck genom vårt paradigm. Vilka dessa aspekter är redogörs för närmare i den fortsatta jämförelsen.

I figur 1 anges plats på blanketten för sådana modifieringar som gäller verbets aspekter. I figur 3 är tempus kodat i kolumn 13 och modus i kolumn 14. Schank (1973, ss 206-207) säger:

"Any conceptualization can be modified by certain <u>conceptual tenses</u> of which ´p ´ for past is one ...
... These tenses modify a conceptualization as a whole."

På samma sätt som vi kodar tema enligt innebörden i satsrelationer, kodar vi tempus och modus med utgångspunkt i verbet. Schank exemplifierar betydelsen av att särskilja relationer från de "conceptual tenses" som ligger i ACT genom meningen: ´Since smoking can kill you, I stopped.´

Den conceptuella representationen är



(c = conditional, t_F = finished transition)

Vad som uttrycks på detta sätt är i den översta satsen: 'man röker' ('smoking'), den underliggande till den: 'man kan bli död (om man röker)' och den tredje under 'smoke' säger: 'jag slutade röka'. Schank kommenterar förhållandet mellan dessa tre conceptualiseringar så här:

"This sentence contains two conceptualizations related by a causal and a causal relating that causal to a third conceptualization. Such a thing is nearly impossible to handle in more traditional linguistic representations."

Detta sista påstående kan inte få stå odiskuterat, så vi skall försöka analysera relationerna enligt ANACONDA-systemet.

De tre conceptualiseringarna 'man röker', 'man kan bli död' och 'jag slutade röka' kan uttryckas genom kodning på datablankett i tre kolumner, som figur 4 visar i det översta exemplet. I Schanks C-diagram ligger en djupare conceptualisering i 'rökning' än vad vi skulle ange. Här har vi för exemplifieringens skull uppdelat begreppet 'rökning' i de två nödvändiga begreppen för en sats. Den andra satsen säger samma sak som Schanks regel 10, alltså förändring av tillstånd från levande till död. Den kunde också ha uttryckts via 30 + 40, dvs 'man kan dö', utan att innebörden hade förändrats. 'Bli' står i vårt fall för "change of state" och resultatet är 'död', som är en beskrivning av actor, således ett beroende begrepp (32). Varianten 30 + 40 innebär ingen skillnad, eftersom den handling som ligger i 'dö' har "sig själv" som resultat. Den tredje satsen har gjorts fullständig genom supplering av den handling som antytts genom verbet 'sluta'. Vi har samma regel här som Schank (1973, s 207) uttrycker med orden:

"The English word 'stop' for example is actually an instance of the conceptual tense 't_F' and thus predicts an ACT. That ACT was unstated ..."

I fråga om tempus har vi inte gjort alla de distinktioner som vore möjliga, t ex början av handling, fortgående handling, avslutande handling, etc. Börja, bruka, sluta osv kodas som verb och tillhör hela verbbegreppet. I exemplet är dessa delade men hör ihop genom samma kod. Ett program finns som sammanför verben till en sträng för vidare bearbetning. De

10–16	25–35	65–67	68–70	71–73
1 1 1	Eftersom	462		
4	(om)		463	
	man			30
	röker			402
	man		30	
	kan bli		41	
	död		321	
2 1 2	slutade	40		
	jag	30		
	(röka)	40		
—	—	—	—	—
1 4 1	Eftersom	462		
4	rökning		30	
	kan döda		40	
	en		501	
2 1 2	slutade	40		
	jag	30		
	(röka)	40		

Figur 4. Kodning av relationer: Modifiering

ovannämnda nyanserna har vi ännu inte tagit ställning till, men det är fullt möjligt att göra en innehållsbestämning i efterhand, när vi har tillräckligt många exempel på hjälpverb eller ofullständiga verb; de skulle då kunna specificeras i kodningsreglerna. Tempus kodas i kolumn 13 med kod 1 för presens, 2 för förfluten tid och 3 för futurum. Om i kolumnerna 13 och 14 står (som i vårt exempel) kod 1 + 4 betyder detta "för talaren närvarande tid, modalt" och 2 + 1 uttyds "för talaren förfluten tid, faktiskt".

Relationen mellan de två första satserna i den översta meningen måste kunna framträda. Schanks c i diagrammet vid den beroende pilen anger "conditional". Vi har angivit denna relation med ett ′om′ (motsvarande "if"). Ett sådant ′om′ blir satsinledare (vilket tidigare nämnts) och får kod 46. Denna relation borde enligt Schank vara en "causal", men på vilket sätt? ′Man kan bli död om man röker′ uttrycker ingen villkorsrelation. Villkoret för att betrakta sig som död är inte att man varit rökare. Villkor måste uttryckas som ett verkligt förhållande. Orsaken till ens död är heller inte rökning, men det kan vara det. Relationen är med andra ord uttryck för en korrelation, men bara för ett potentiellt orsakssamband. Vi kodar inget tema för just den relationen, men vi anger samband genom

kod 46. I stället uttrycker vi modalitet genom kod 4 i kolumn 14. Detta samband, den potentiella orsaken, hänger sedan ihop med att 'jag slutade röka'. Satsinledare för detta är 'eftersom', som följdriktigt har kod 46. Innebörden i hela meningen skall tydas: 'Eftersom det finns ett potentiellt orsakssamband mellan rökning och död, slutade jag röka.' Detta är orsaken till att jag slutade röka och detta senare orsakssamband måste uttryckas genom en temakod, nämligen i kolumn 16. Att jag slutade röka är en följd (2) av att jag har funderat över riskerna (1).

Av framställningen har det förhoppningsvis framgått att kod 46 pekar mot en sats, framåt eller bakåt. Detta anges med siffra för respektive sats ordningsföljd i meningen. I dator går läsningsprocessen till så att varje begrepp (kort) läses för sig i den ordningsföljd de står. Det kan nu vara så, att vi i våra analyser vill arbeta med de enskilda begreppen utan identifikation, tema eller plats i meningen, genom kopplingar av begrepp. Loop-koden är till för att begreppskoderna inte skall "hänga i luften". För att exemplifiera loop-koden kan vi göra följande sökning. Vi vill veta hur olika verb står i förhållande till varandra (vi räknar 41 + 32 som tillståndsverb) och maskinen samlar tre verb ur denna mening i den följd de står. Några andra begrepp har vi inte att gå efter. Det första verbet är kodat i kolumnerna 72 och 73, vilket framgår av utskriften (kortbilden). Då undrar vi: Hur hänger verben ihop? Maskinen får en instruktion i kolumn 74, som säger "gå till den andra satsen". Där träffar den på kod 46. I den satsen finns det andra verbet. Då kan vi sluta oss till att det andra verbet är överordnat det första. Efter det andra kommer det en instruktion att gå till sats nr 1. Där ger en ny kod 46 ett samband med det tredje verbet. Utan att "veta" meningstillhörighet kan datorn på detta sätt gruppera begrepp som bildar en struktur.

Nu går vi tillbaka till det exempel som Schank utgick ifrån: 'Since smoking can kill you, I stopped.' 'Smoking' var detsamma som 'man röker'. Det första exemplets tredje sats dras alltså ihop till ett begrepp 'rökning'. Då är vi över i sats två genom looping. 'Kan bli död' var detsamma som 'kan dö'. Hur hänger 'man kan dö' ihop med 'rökning'? Jo, looping säger att det kan finnas ett orsakssamband: 'man kan dö på grund av rökning', alltså 'rökning kan döda en'. Exempel 2 i figur 4 uttrycker detta. Vi ser då begreppet 'rökning' som agent och 'man' som offret ('en' = objektsform). Ingenting i conceptualiseringen har egentligen förändrats; temakodningen är densamma. Som det sista exemplet visar skulle ANACONDA ha representerat meningen. Att rökning innebär 'man inhalerar rök genom att en cigarrett (eller hellre rökverk) transporterar röken från sig själv till 'man' anser vi oss inte behöva koda. Hur man i ett lexikon skall specificera begreppet rökning semantiskt är en långt senare fråga.

Här har vi försökt visa att ANACONDA, trots sitt relativt enkelt uppbyggda kodsystem,ändå kan handha komplexa relationer. Ett påpekande återstår dock att göra till de här diskuterade kodningarna. Vi får inte fram det faktum, som ligger implicit i 'jag slutade röka', nämligen 'jag rökte (eller har rökt)', vilket alltså innebär att jag har 'börjat röka och fortgående hållit på att röka, samt gjort det med vissa mellanrum'. Men det gör å andra sidan inte Schank heller.

De andra conceptuella relationerna enligt Schank är tid ("time") och plats ("location"). Två regler symboliserar tidsrelation, där den ena avser ett begrepp och den andra anger att en conceptualisering är en tidsaspekt av en annan conceptualisering. Den tredje indikerar att en händelse måste äga rum någonstans. I ruta 11 symboliseras dessa tre regler tillsammans med motsvarande ANACONDA-koder.

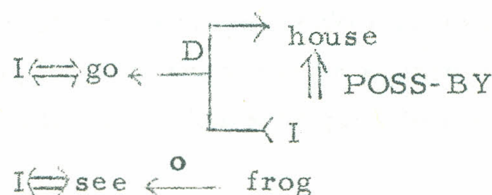
Ruta 11. Jämförelse mellan C-diagram och ANACONDA: Tids- och platsrelationer

C-diagram		ANACONDA
Regel 12.	T ↓ ⟺	43
Regel 13.	⟺ ↓ ⟺	43 (+ loop + sats)
Regel 14.	PP ⇓ ⟺	44

"T" i regel 12 står för ett tidsbegrepp, som t ex 'i går' 'klockan 12'. Begreppet relateras till en hel conceptualisering. Schank (1973, s 207) menar: "The time of something modifies the entire conceptualization and not any particular item in it."

Liksom i fallet med plats refererar tid inte till ACT, såsom ett kasus, som måste finnas explicit till vissa verb. Vi har av praktiska skäl anpassat dessa koder till traditionella beteckningar av dessa begrepp, nämligen tids- och rumsadverbial, som är lättast att känna igen vid satsanalys. Koderna anger därför ett beroende till verbet. Eftersom verbet är det centrala i en conceptualisering är detta den lämpligaste kodbeteckningen, ifall man inte vill skapa ytterligare koder med ny första siffra.

Regel 13 förklaras med exemplet 'While going home I saw a frog', som representeras



Tecknet för tid uttrycker två handlingar, som egentligen är svåra att särskilja i tiden, så att den ena föregår den andra (jfr diskussionen till fig 3). Den översta meningen uttrycker tid, innefattande en riktnings mål, och den undre betecknar handlingen. Tidssatsen skulle kunna dras ihop till 'på hemvägen'. Tidsadverbet är i det här fallet på svenska 'medan', 'under tiden som' och inte 'när'. Vår kodning enligt regel 12 och 13 visas i figur 5.

25–40	45–71
Medan	432
jag	30
gick	40
hem	441
såg	40
jag	30
en groda	50
på hemvägen	43
såg	40
jag	30
en groda	50

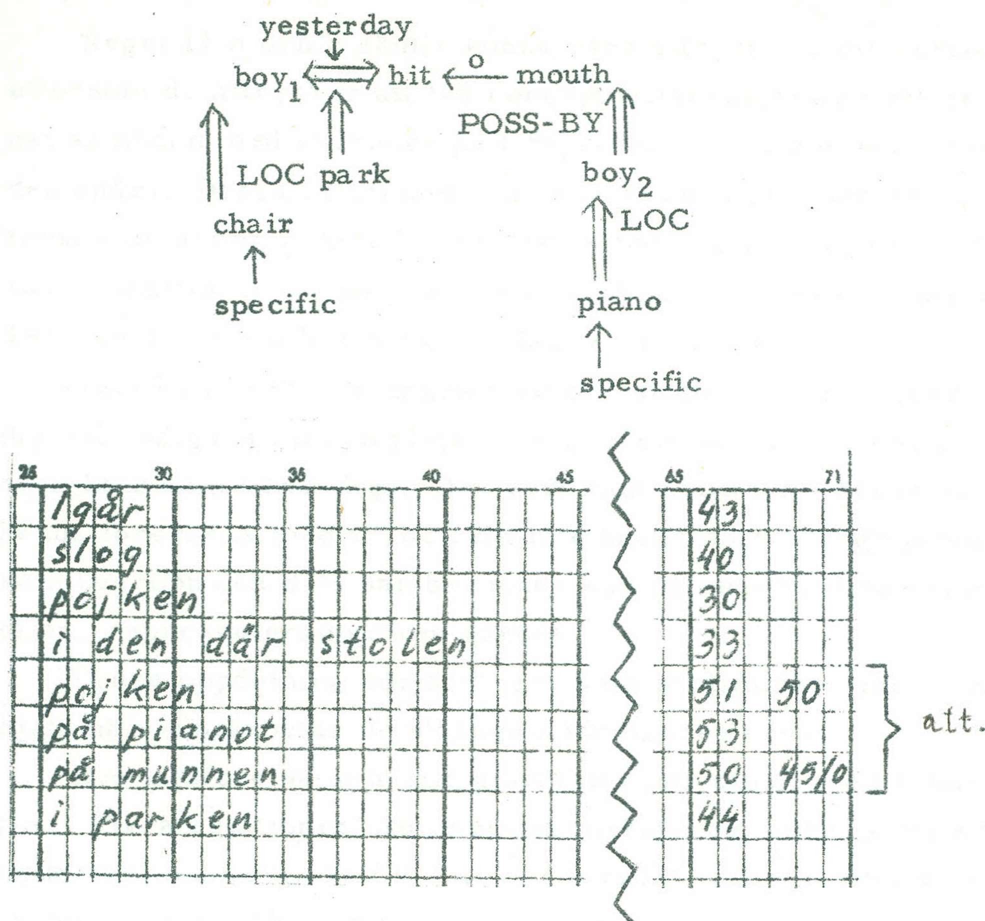
Figur 5. Kodning av tidsbegrepp och -relation

Det översta exemplet är kodat enligt regel 13 och följer Schanks representation. I regel 12 (ruta 11) ser vi att pilen för tid inte är dubbel, dvs inte skulle innehålla ett PP. Vårt 'hemväg' skulle heller inte kunna tolkas som en konkret väg genom koden för tid, så därför skulle det inte innebära någon risk att dra ihop satsen till ett begrepp.

Regel 14 är av intresse, eftersom det finns olika slag av platsbestämning. Denna regel avser inte sådan bestämning som är beroende till ett PP (enligt regel 5, ruta 5). Schank ger ett intressant exempel på dessa skillnader och det skall återges här med ANACONDA:s kodning i direkt anslutning. Meningen lyder:

'Yesterday, the boy in that chair hit the boy on the piano in the mouth in the park.'

Vi ser av figur 6 att de båda LOC är beroende PP och de kodas med vertikal beroendekod (33, 53). 'Park' anges mellan agenten och handlingen och kan därför ses som beroende till den händelse som utspelades; 'i parken' kodas med 44. Så långt är meningen enkel att dela upp på vårt sätt. Anledningen till att exemplet är intressant är kodningen av 'på munnen', som trots prepositionen inte kan tydas som platsbestämning av något slag.



Figur 6. Teoretisk och praktisk representation av lokalitetsbegrepp

Som Schank anger är pojke nr 2 ägare av munnen som pojke nr 1 slog på. Då skall ´munnen´ kodas som huvudord och ´pojken´ som bestämning i genitiv, enligt vårt första förslag. Hela begreppskomplexet uttyds då ´pojkens på pianot mun´, vilket torde vara det rätta. Men egentligen säger vår kodning inte exakt det. Kod 53 skall vara en utfyllnad till huvudordet i komplexet och då skulle vi få ´på munnen på pianot´. I sådana fall blir en sökning utvidgad så att när kopplingen mellan två begrepp inte är semantiskt meningsfull utvidgas sökningen till att gälla även kod 51, där beroendeförhållandet framträder. Det andra exemplet innebär att vi inte anser ´munnen´ vara huvudsaken, utan att hela pojken blev slagen. Var slaget träffade är då en upplysning som har med handlingen att göra, dvs handlingen i ´slog´ är ´rörde häftigt handen (?), så att den träffade munnen´ och är det sätt på vilket det hela gick till, således kod 45. En tredje möjlighet anges med /0 och som skall förstås så att vi t o m kan tänka oss att ange hela handlingen (40 + 50) som ACT, dvs kod 40, som då ses som det idiomliknande uttrycket ´slog på käften´, där ´käften´ inte alls skall förstås bokstavligt. De två sista alternativen tas upp här för att visa kodsystemets plasticitet. Den första kodningen är troligtvis den mest adekvata och följer Schanks modell.

Regel 11 a och b skulle kunna vara uttryck för ett horisontellt beroende, eftersom de uttrycker att två conceptualiseringar har ett orsaks- eller annat samband med varandra på tidsplanet, så att den ena i tiden förutsätter den andra. Dessa relationer kodas i vårt system med 46 + loop samt med tema som anger vilket slag av bisats meningen innehåller. Bisatsen kan vara konditional, kausal, koncessiv, konsekutiv eller final (kolumnerna 15-18). (De två sista har sammanslagits, se sid 42.)

Vertikalt beroende anges med de relationer som ligger utanför paradigmet, således tid och plats, enligt reglerna 12-14, där regel 12 och 13 betraktas som likvärdiga. Den vertikala aspekten i dessa relationer skulle kunna försvaras med att de dels inte kodas som övergripande tema utan som begrepp och dels har beroende kod till den händelse som meningen omtalar, representerad genom verbet.

Denna uppdelning bör dock ses som preliminär, men kan vara till hjälp när vi utarbetar de slutliga kodningsreglerna.

Den jämförelse som har gjorts här har utgått från Schanks presentation. Flera relationer skulle kunna diskuteras, men lämnas åt sidan i detta sammanhang. Vad vi här har försökt visa i kapitel 5 är på vilket sätt en teori om språk i form av symboler och bakomliggande tankestrukturer kan representeras i form av siffror för input i dator. Grundläggande är de 14 reglerna för begrepp och relationer. I sin fortsatta framställning går Schank över på semantiska spörsmål ("conceptual semantics"). Vi är ännu inte beredda att gå in på en jämförelse därvidlag av det skälet att vi baserar "innehållet" i begreppen på en empirisk utprövning (se kap 3.4), där de statistiska analyserna inte är färdiga.

6. ANALYS AV TANKESTRUKTURER

I diskussionen om General Inquirer framhölls vikten av att bygga upp lexikon på grundval av en språkteori. Målet med den teori om språk som vi följer är endast en representation av språkstruktur. Det har framgått att vi inte alltid anser oss behöva koda texten, eller representera alla underliggande conceptualiseringar. Det beror dels på att vår manuella bearbetning av texten skulle bli för tidsödande och kanske också innebära svårigheter vid samstämmighet bland flera kodare, som är en viktig faktor när man skall handha stora mängder av text. Dels och främst beror det på att alla underliggande strukturer inte är nödvändiga för vårt syfte, för den information vi önskar få ut ur materialet. En teori för kodning av verbala data måste anpassas till en teori om den "värld" som dessa data är en representation av, i vårt fall pedagogiska forskares syn på sin egen värld. Denna värld ses som ett system av värderingar och kunskap, vilket bestämmer vilka handlingar den enskilde utför.

Sedd i detta perspektiv ansluter sig vår analys till kognitionspsykologisk forskning. Inom ramen för denna forskningsgren har Abelson (1973) utvecklat en modell för analys av "belief systems" (belief = ungefär försanthållande, övertygelse). I "The structure of belief systems" anger Abelson förutsättningarna för utvecklingen av en "ideologimaskin", en simulator, som konstruerar möjliga politiska syften, planer, aktioner och tillstånd (t ex i "det kalla kriget") med utgångspunkt i samspelet mellan olika politiska övertygelsesystem. Många av Abelsons idéer har varit betydande för GI utveckling. Detta är ett skäl till att Abelson även för oss fått stor betydelse. En väsentligare orsak är emellertid att hans framställning (Abelson, 1973) grundas på Schanks språkteori i modifierad form samt att hans modell för analys har en målsättning som i många stycken liknar vår. Den följande framställningen skall därför presentera Abelsons system och diskutera dess relevans för ANACONDA.

6.1 Teoretisk ram

Inledningsvis anför Abelson (1973, s 293) följande:

"In the process, we make extended use of Schank's Conceptual Dependency Analysis. ... It is essential in giving firm conceptual grounding at the level of event detail. But our emphasis is different from Schank's in that we are oriented more toward the relationships between conceptual dependencies than toward representations within conceptual dependencies. Consequently, our notation is occasionally at variance with Schank's, mainly in the service of convenience rather than because of theoretical disagreement."

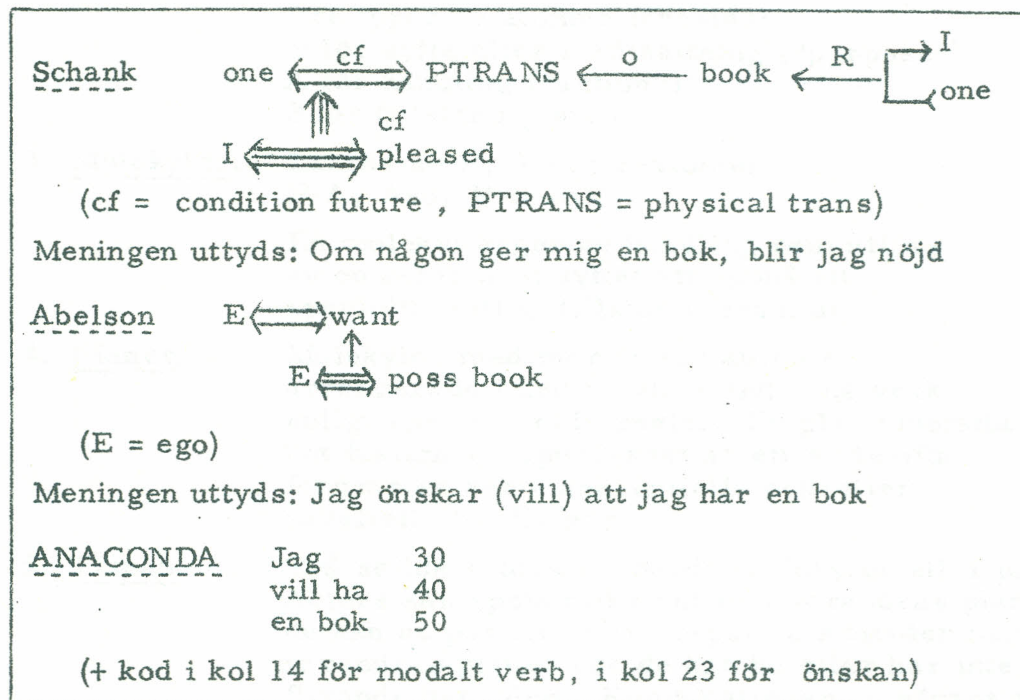
Därefter nämns några skillnader, som avser djupet i representationen, där det för oss väsentliga ställningstagandet gäller vissa verb. Vi har

redan nämnt att Schanks regel 10 inte är relevant, eftersom den innebär kodning av ett resultat som ligger implicit i verbet. Abelson (1973, s 294) exemplifierar en av skillnaderna med verbet 'want':

"We emphasize the directed activity to which a 'want' can give rise, rather than the state of pleasure which ensues from a satisfied 'want'."

Meningen 'I want a book' skulle analyseras olika hos Schank och Abelson. Jämförelsen görs i ruta 12 tillsammans med vår kodning.

Ruta 12. Jämförande analys av verbet 'want'



Som ruta 12 visar är den huvudsakliga betydelsen av verbet 'want' hos Schank det tillstånd som uppfyllandet av önskningen ger. Abelson däremot anger att målet för önskningen är att komma i besittning av boken.

Överhuvudtaget finns det ett problem med sådana här jämförelser. De exempel som ges är inte satta i ett sammanhang, dvs man har inte kontexten klar för sig. Detta faktum aktualiseras i exemplet på vår kodning här. Om meningen uttrycker att 'jag vill ha en bok för att det är lämpligt att ha en när jag ska göra det här', är detta inte samma sak som 'jag önskar att jag hade en bok (eftersom jag inte har någon alls)'. I det första fallet anger vi ingenting i kolumn 23. "Önskan" gäller när det är någonting man önskar få eller uppnå, som man hittills inte har. Våra regler ger således utrymme för båda tolkningarna av 'want': 'brukar vilja ha' och 'skulle vilja uppnå'. Vi ansluter oss dock närmast till Abelson, eftersom 'ha' är huvudsaken.

Det finns både likheter och skillnader mellan Abelsons föreslagna system och vårt och det har varit intressant att ta del av de strukturer som

Abelson föreslår finnas i ett "belief system". Denna hierarki presenteras nedan och därefter följer en exemplifiering av de olika begreppen.

Abelson (1973, ss 294-295) diskuterar med utgångspunkt i följande nivåer:

1. <u>Element.</u> De minsta enheterna, basen i ett lexikon

2. <u>Atomer.</u> Enkla strukturer av element, som länkas ihop i ett C-diagram

   Tre typer av atomer föreslås:
   P för syfte eller förutsättning ("purpose")
   A för handling ("action")
   S för tillstånd ("state")

3. <u>Molekyler.</u> Länkar av P, A och S-atomer; P-A, A-S, P-S-par

   En molekyl anger en handling som utförs av en actor med syftet att uppnå ett speciellt slutligt tillstånd (resultat)

4. <u>Planer.</u> Molekyler med mer än tre atomer, arrangerade i kedjor eller andra nätverk enligt specificerade regler. En plan innefattar det faktum att uppnåendet av ett syfte ofta fordrar en serie sekventiella och/eller parallella handlingar

5. <u>Teman.</u> Två skilda actors beroende molekyler eller planer. Actors kan spela olika roller i varandras planer; de kan ha positiva eller negativa attityder mot varandras planer och de kan ha eller har inte inflytande över dem. Kombinationer av sådana möjligheter ger upphov till en taxonomi av teman.

6. <u>Scripts.</u> Sekvens av teman som involverar samma slag av actors, med förändring i beroenden ("interdependencies") från ett tema till ett annat; utveckling av en "story" av möjliga förändringar i relationer mellan actors.

6.2 <u>Kodningsreglers tillämpning i en hierarkisk struktur</u>

De elementstrukturer som Abelson kallar atomer motsvarar vad Schank kallar conceptualization, där elementen motsvaras av begrepp. De består alltså av actor, action och de "kasus" som är nödvändiga för att specificera handlingen. Abelson (1973, ss 295-297) gör en indelning av atomerna i tre typer, vilka symboliskt presenteras på följande sätt (ss 295-297):

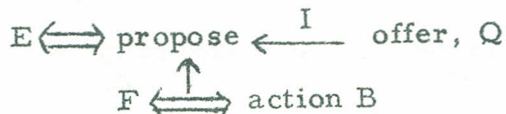
<u>A (action) atoms</u>   one ⟺ trans $\xleftarrow{O}$ X $\xleftarrow{R}$ [→ E ; → one] $\xleftarrow{I}$ Y

Meningen uttrycker att någon ger objektet X till Ego, som är recipient. Y anger möjligheten av att ett instrument är nödvändigt för transfereringen.



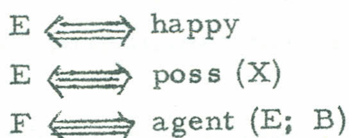
Detta diagram uttrycker "directional case", där ett objekt flyttas från plats α till plats β.

En mycket viktig handlingsatom i Abelsons system är den där handlingen är ett förslag från en actor till en annan actor:



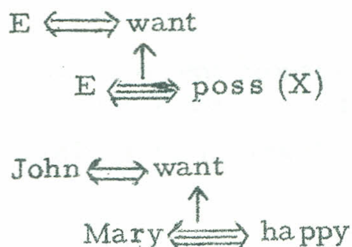
Meningen uttyds: 'E föreslår genom ett erbjudande av något slag att F skall utföra handlingen B.'

S (state) atoms



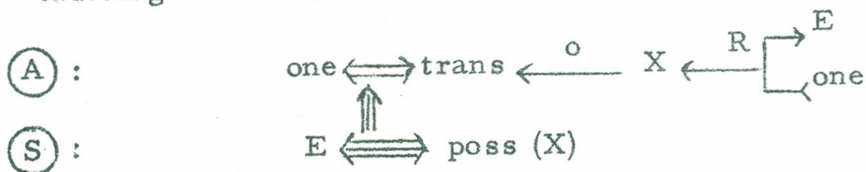
Genom de tre pilarna uttrycker Abelson (1973, s 296) ett tillståndsberoende, som är ett resultat hos en actor eller sak, antingen ensamt eller i relation till ett annat objekt eller sak. "Agent" har särskild betydelse i detta system. När F accepterar förslaget av E att utföra handlingen B, blir F agent till E. Distinktionen mellan actor och agent är väsentlig i ett system, som skall spegla samspelsstrukturer med avseende på t ex sociala kontakter och politiska beroenden.

P (purpose) atoms

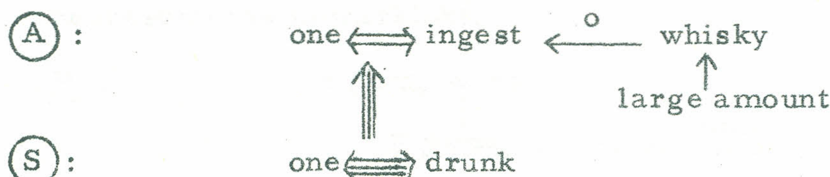


En P-atom uttrycker en önskan tillsammans med det tillstånd som önskas. Den nedre atomens actor behöver inte vara densamma som den övre atomens.

A- och S-atomer kan vara förbundna med varandra om handlingen i A-atomen sägs vara orsaksrelaterad ("causal relation") till S-atomen. Abelson (1973, s 297) beskriver några typiska exempel, varav den s k "tautological causation" är den vanligaste:

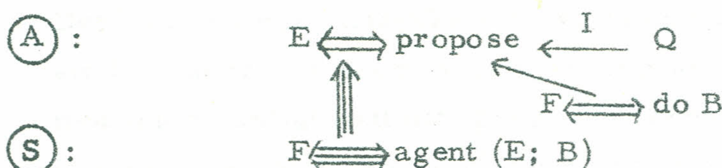


Denna bindning ("bond") skulle kunna översättas med: "Om någon överför X till E, så blir E ägare av X". En tautologisk orsak är således det tillstånd (S) som handlingen med nödvändighet resulterar i. Andra bindningar innebär naturliga konsekvenser av användningen av vissa objekt, där det enda villkoret för bindningen är att en handling har tillräcklig styrka eller kvantitet, t ex:



som skulle kunna översättas med: "Om man dricker en stor mängd whisky, så blir man full." Sådana paradigmer kallas av Abelson "conceptual postulates".

Det tredje exemplet är mer komplext och framställs så här:



Detta något komplicerade diagram ställer frågan: /Vad är det som gör att F blir agent till E och utför handlingen B, efter att E har föreslagit det till F med belöningen Q?/ Svaret kan enligt Abelson (1973, s 298) generellt vara detta:

"... F will accept when the apparent positive value of Q to him is greater than the apparent negative value of performing behavior B."

Tre atomer, P, A, S bildar en molekyl när tre villkor är uppfyllda, ett mellan varje atompar:

(PS). S-atomen är det tillstånd som förbindes med P-atomens önskan.

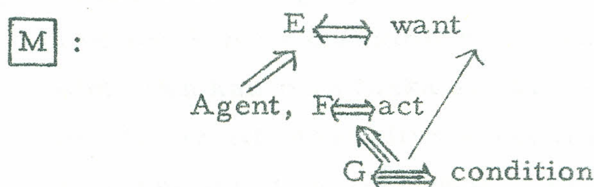
(AS). A-atomen är relaterad till S-atomen och

(PA). Actor i A-atomen är agent till Actor i P-atomen.

Om molekylen som enhet säger Abelson (1973, s 299):

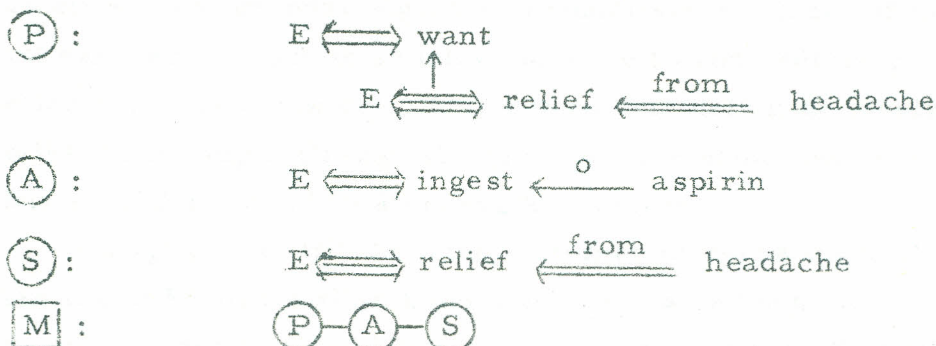
"The molecular unit captures the idea of an action undertaken in order to attain a goal desired by the sponsor of the action. It is the essential building block of all belief systems which find meaning in the purposeful activities of individuals, institutions, and governments - or even the animistic forces of nature or gods."

En molekyl kan representeras genom ett C-diagram:



De tre horisontella linjerna (eller atomerna) representerar "motivation", "action" respektive "goal". De diagonala enkla, dubbla respektive fyllda pilarna anger "purpose", "agency" och "causation". "Actor" behöver inte vara densamma i de tre atomerna, men kan vara det. Flera möjligheter finns, t ex EEE, EEF, EFE, EFF, EFG.

En kedja av atomer som resulterar i en fullständig P-A-S molekyl kan representeras hierarkiskt:



Det innehåll som uttrycks i denna hierarki är ungefär: E vill bli av med sin huvudvärk. E äter aspirin. E blir av med huvudvärken. Detta är däremot inget vanligt sätt att formulera innehållet i denna molekyl, snarare ett ovanligt klumpigt sätt. Det låter bättre om man säger 'E ville bli av med sin huvudvärk, så han åt aspirin och så blev han av med den.' Men i realiteten, på den syntaktiska nivån, uttrycker man hellre denna innebörd på ett mera ekonomiskt sätt, t ex: 'För att E skulle bli av med sin huvudvärk åt han aspirin' eller 'E åt aspirin (så att han skulle) för att bli av med sin huvudvärk' eller 'Genom att äta aspirin blev E av med sin huvudvärk'. Att huvudvärken gick bort antas, eftersom ytterligare försök att bli av med den inte sagts. S-atomen ligger alltså implicit i P-A-relationen.

Efter denna genomgång av de grundläggande enheterna för Abelsons analys skall dessa diskuteras i förhållande till ANACONDA.

Vi har ovan konstaterat att en atom motsvaras av en conceptualisering, alltså ett fullständigt ANACONDA-paradigm i form av en mening eller sats. En molekyl motsvaras i sin enklaste form (P-A-S) av relationer, dvs satser som är beroende av varandra.

De två först exemplifierade atomerna är likadana som de conceptualiseringar vi redan presenterat kodningen av. Till det andra exemplet kan dock tilläggas att, eftersom vi ännu inte har skilt på plats och riktning, vi kodar dessa begrepp med kod 44 i båda fallen (30 + 40 + 50 + 44 + 44). Liksom Schank diskuterar Abelson instrumentets roll. I fallet vid exemplet ovan kan man tänka sig att handlingen utgörs av verbet 'skicka', som innefattar ett visst eller vissa instrument eller medel för handlingens genomförande, t ex 'per post'. Vissa förutsättningar måste vara uppfyllda

för att "transfer" kan genomföras, något som Abelson benämner "instrumental control". Om instrumentet inte finns explicit, kodas det inte hos oss.

Ägande, plats och innehåll ("possession", "location", "containment") som hos oss motsvaras av kodernas andra siffra i begreppskomplexet (31, 33) är vanligtvis inte nödvändiga i Abelsons system. Till detta ansluter vi oss, eftersom vi i första hand arbetar med huvudbegrepp.

Det tredje exemplet är också det första exemplet på hur uppbyggnaden av ett system grundar sig på en samhällsteori. Man måste hålla för sant att när flera personer är inblandade i ett händelseförlopp, vissa påverkar eller styr andra och det bildas beroenden mellan de involverade, ensidigt eller ömsesidigt. Frågan för oss i det här sammanhanget är tills vidare hur en sådan atom kan återspeglas i texten.

I diagrammet till den tredje atomen uttrycks ingen orsaksrelation mellan E och F, dvs mellan handlingen "propose" och "action" B, bara att B är underordnad i tiden till den förra. Dessutom är E endast initiativtagaren och utför inte själv en handling som skall relateras till den senare. Enklast kan man uttrycka detta med 'E föreslog att F ...'. En att-sats betraktas som ett begrepp, bestående av flera begrepp, alltså E (30) föreslog (40) att ... (50), som egentligen inte innebär något annat än en conceptualisering. De begrepp som innefattas i att-satsen står samtliga i objektsförhållande till verbet. Detta exemplifieras i figur 7.

25 30 35 40 45	65 71
Professor X	30
uppmanade	40
mig	70
att	50 2
(jag)	30
(skulle) söka	40
tjänsten (Y)	50

Biblioteket	30
krävde	40
att	50 2
jag	30
skulle återlämna	40
böcker	50
(till biblioteket)	70

Figur 7. Kodning av handlingsatomer: Att-satser

Handlingarna ´söka´ respektive ´återlämna´ följer i tiden efter ´uppmana´ och ´kräva´. ´Jag´ är i båda fallen ett objekt för professorns och bibliotekets handlingar. Här framgår inte om ´jag´ agerar enligt intentionerna. Vad som visas här är att det kunde vara möjligt att genom kodning av en enda mening få fram vad Abelson menar med "actor" och "agent". Här framgår det dock inte om agenten ´jag´ verkligen utför handlingarna. Vad som återspeglas i denna struktur är möjligen vem som är "actor", dvs överordnad och det anges med huvudsats. Vem som i realiteten är agent till vem avspeglas sällan inom en mening (eller atom). Det ligger förborgat inom ett helt "system" och bör diskuteras på en annan nivå, om diskussionen alls är möjlig för oss.

När det gäller S-atomerna motsvaras den första av det vertikala beroendet mellan begrepp, t ex 30 + 32, som har att göra med kopulaverb (eller predikatsfyllnad enligt traditionell satsanalys).

Den andra som uttrycker ägande betyder inte detsamma som Schanks POSS eller vår kod 31. Som den efterföljande texten anger är S ett tillstånd som är en conceptualisering, dvs har något slag av verb som länk mellan minst två begrepp. I det här fallet skulle verbet vara ´ha´ (äga). I den text som är kodad enligt vårt system har verbet ´ha´ inte brutits ur de övriga verben, men det har diskuterats och det kan mycket väl göras i efterhand i och med att trunkeringar gjorts på samtliga verb. Vi kan alltså maskinellt ta ut alla former av ´ha´ och byta ut kod 40 mot en ny kod. Frågan är bara om detta är nödvändigt på kodningsstadiet. Det är riktigare att betrakta verbet ´ha´ som bärare av ett tillstånd, men det begrepp som följer är inte ett beroende (adjektiv) utan ett oberoende, så någon skillnad borde i så fall anges i kodningen av verbet. De olika verben skall dock i ett senare skede specificeras och kategoriseras och det torde finnas flera som bedöms vara av samma slag som ´ha´ i fråga om Osgoods tre dimensioner, t ex s k statiska verb. Ett motiv för att ändå särskilja ´ha´ från andra kunde vara den höga frekvensen det har.

Liksom det slag av A-atom som diskuterades i ruta 12 kan den tredje S-atomen knappast i sin helhet diskuteras som en urskiljbar företeelse på det syntaktiska planet. Visserligen kan beroendet mellan F och agent uttryckas genom koderna 30 + 30/33, men det förutsätter att detta sägs explicit (F är agent till E). Att förhållandet eller tillståndet existerar kan bara analyseras fram genom att man tar hänsyn till handlingar som F utför och som visar att de är önskade av E. För att kunna fastställa ett sådant tillstånd i systemet måste man gå utanför meningsplanet ("sentential level"). Det är mot denna bakgrund man får betrakta de sista av Abelsons exempel på A-S-par. Detsamma gäller vad "Q" står för i det enskilda fallet.

De två första paren är generellt viktiga i ett system för försanthållanden. De är båda impliceringar, som var och en som har förstått handlingarna håller för sant. Sådana conceptuella postulat kodar vi inte av samma skäl som vi inte kodar Schanks implicita instrument.

I anslutning till P-A-S-molekylen angavs ett antal alternativa meningar som denna molekyl kunde ha uttryckts i på ett syntaktiskt plan. Meningarna kodas då antingen genom att två satser relateras till varandra, såsom de sista exemplen lydde, eller som något av de första, där de tre satserna till formen är huvudsatser. Om vi här utgår ifrån att S-atomen fortfarande inte är relevant kunde en kodning av ett par alternativ se ut som figuren nedan visar.

men. nr					
0001	01	1	E		30
	02		åt		40
	03		aspirin		50
	04	2	för att	462	
	05		(E)		30
	06		(skulle) bli av med		40
	07		huvudvärken		50
0001	01	1	E	30	
	02		ville bli av med	40	
	03		huvudvärken	50	
0002	01		så	99	
	02		han (E)	30	
	03		åt	40	
	04		aspirin	50	

Figur 8. Exempel på kodning av en enkel molekyl

Det första vi kan konstatera i dessa två kodningar är att båda är uttryck för samma innehåll. Det första fallet är en syntaktisk mening med en bisats som uttrycker följd eller avsikt. Det tillstånd som finns implicit i bisatsen ligger i tiden efter huvudsatsens handling. Om vi översätter Abelsons beteckningar på denna kodning får vi molekylen A + P(S).

Det andra består av två meningar, eftersom de är två huvudsatser. Den första uttrycker en önskan med samma implicita resultat som det första exemplets bisats. Men i Abelsons resonemang måste en molekyl bestå av ett par. Vi måste alltså gå över i nästa mening (se kolumn 8) för

att få komplementet, som är den handling som erfordras. Vi får då mönstret P(S), A.

Detta aktualiserar problemet med syntaktisk kodning, framför allt avgränsning av mening respektive sats. En syntaktiskt fullständig mening, utgående från begreppen huvudsats och bisats, har vi ansett vara det säkraste sättet att koda texten i enheter. Om den formella strukturen i kodningen är specificerad blir det betydligt enklare att analysera relationer. Det senare exemplet i figuren skulle också kunna kodas i kolumn 18, övergripande två meningar. Men det blir betydligt svårare om P-A och S inte finns i omedelbar närhet av varandra. Vår princip med satskolumner och looping går av utrymmesskäl inte att genomföra om de sammanhängande satserna blir fler än fem. Att gå ut över meningsplanet innebär dessutom en tolkning av beroendeförhållanden, vilket i så liten utsträckning som möjligt skall ske på det stadiet. Det är således stor skillnad mellan ett fullständigt paradigm och en fullständig molekyl. Paradigmet kodas enligt de regler vi anfört, men molekyler, dvs relationer mellan meningar är en struktur som vi avser analysera via en datorprocess. För att kunna göra det måste vi veta vad de olika verben har för innebörd, varvid en serie verb av ospecificerat antal kommer att höra ihop. Denna princip blir tydlig när man har insett vad Abelson menar med den näst högre nivån, planen.

Abelson (1973, s 300) skriver:

"... if intermediate steps toward the goal are required, then the molecule must be larger. /.../ Whether several necessary steps in achieving a purpose are obvious all along or whether they are originally unsuspected but later revealed, we refer to the enlarged molecule as a 'plan'".

Man kan exemplifiera innebörden i detta med att E, som ville ha aspirin, inte hade det hemma, utan först måste gå till ett apotek, som var stängt. Han går då till nästa apotek, osv. När han kommer hem kanske vattnet är avstängt, så han kan inte ta sin tablett, eftersom han aldrig har kunnat svälja tabletter utan vatten. Många handlingar av en eller flera personer kan föregå uppfyllandet av E:s uppställda mål. När målet inte kan uppfyllas direkt måste en planering ske, så att kopplingen av flera handlingar och tillstånd till slut leder till målet. Det kan således vara ett helt nätverk av "delmål".

En plan kan vara serial eller trädlik. Den förra typen kallas för kedja, som enligt Abelson (1973, ss 305-306) måste uppfylla följande villkor

a) en slutlig S-atom i kedjan är det tillstånd som förbinds med ett 'want' i P

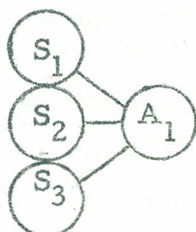
b) varje A är relaterad till en efterföljande S

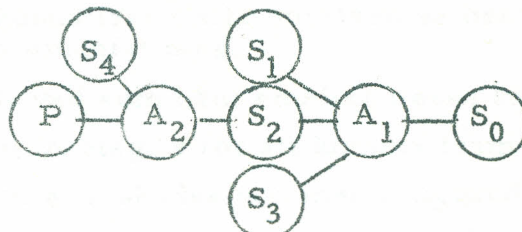
c) actor i varje A är antingen actor i den initiala P eller agent för actor i den följande A (eller den initiala P om ingen föregående A finns)

d) varje S "möjliggör" den A-atom som följer

[M] : 

När flera S är nödvändiga för en A kallas det träd:

[M] : 

eller 

Det senare trädet representerar en plan i vilken actor utför A_2 (med hjälp av det utanförliggande tillståndet S_4) för att producera S_2, som (tillsammans med de utanförliggande tillstånden S_1, S_3) möjliggör action A_1, som medför resultatet S_0.

Dessa utvidgade molekyler för diskussionen upp på en nivå, som inte längre berörs av kodningsreglerna i vårt system, eftersom det nu gäller kopplingar mellan molekyler, dvs en serie kodade meningar av skiftande längd. Dessutom är det inte säkert att en kedja eller ett träd återspeglas i den syntaktiska strukturen. Detta torde stå klart i fråga om det sista trädet, där de utanförliggande S-atomerna symboliserar förutsättningar som kan ligga implicit i handlingars utförande. Så kan t ex förutsättningen för att en forskare i vår intervju använder sig av ett visst informationsorgan vara att han vet hur han skall använda det, att han har gjort en sökprofil för sökning i ERIC-systemet, osv. Om sådana uppgifter finns, kodas de självfallet, men i annat fall kodar vi bara uppgiften om att han använder det eller det organet.

Kärnan i en plan är verben, som för handlingen framåt mot ett mål (i bästa fall). Den information som en person ger om ett händelseförlopp är den som han själv anser vara nödvändig för att lyssnaren skall uppfatta förloppet. Abelson (1973, s 310) skriver:

"The ... verbs refer to the actions..., the heart (or should one say spine?) of the plan. The rest of the structure has to do with the enabling knowledges and accesses, the little things which must be true for the plan to succeed, but which are often left implicit in informal discussion."

Detta resonemang är alltså av samma slag som Schanks. "The little things" hänför sig till de conceptuella reglerna, som en kognitiv struktur består av. För att få en uppfattning om i vilken utsträckning en plan är skönjbar på den syntaktiska nivån har vi gjort en sökning i vårt material, där vi specificerade följande: Skriv ut samtliga meningar som har kod (1, 2) i kolumn 18. Som redan visats innebär denna kodning att vi får uppgifter om en önskan eller avsikt tillsammans med en handling som är relaterad till målet. Vi skall nedan exemplifiera utskrift från en ip. (Frågan gäller information och hur den eventuellt kan förbättras.) Den första meningen har nr 35 och den sista nr 41, som var de två som återvanns:

35. "Det är naturligtvis bra att få så mycket som möjligt av översikterna gjorda så att man vet vad man själv ska läsa."

41. "Det finns inte en person som skulle kunna täcka alla områden så bra att han kan klara av att göra en god sammanfattning."

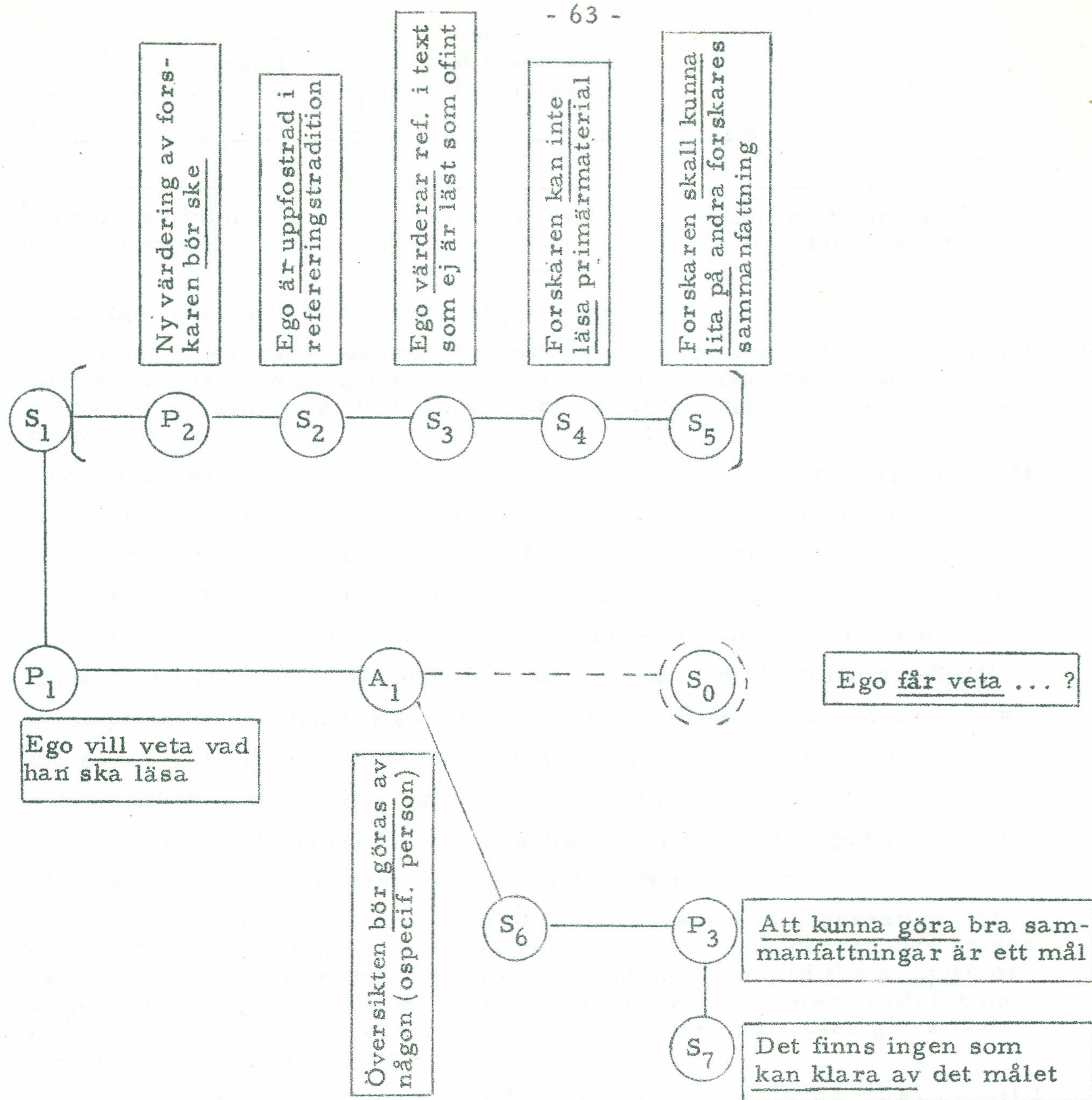
Det är otillfredsställande att inte känna till vad som hänt emellan dessa två meningar, att inte veta vilken strategi som företagits för att komma fram till vad den sista meningen uttrycker. Detta granskades och mellanliggande meningar kompletterades. Vi har försökt ange dem enligt symbolerna för en plan, som ruta 13 nedan visar.

Ruta 13. Sökning i ANACONDA-materialet: Meningar enligt en plan

Men nr	Text	Symbol
35	Det är naturligtvis bra att få så mycket som möjligt av översikterna gjorda så att man vet vad man själv skall läsa.	A + P
36	Man måste komma till en annan värdering av forskaren	P
37	Jag är uppfostrad i den traditionen att referera	S
38	Att ha referens i texten som man inte har läst är ju mycket ofint	S
39	Det finns ingen chans för forskaren att sitta och läsa primärmaterialet	S
40	utan han skall mycket väl lita på vad andra har gjort	A
41	Det finns inte en person som skulle kunna täcka alla områden så bra att han kan klara av att göra en god sammanfattning	S + P

Utgående från den första meningen kan vi säga att forskaren ifråga (Ego) önskar att han fick veta vad det är som just han ska läsa för att ha den bästa tänkbara informationen om sitt eget område. Det slutliga tillståndet skulle då vara att han har tillgång till det han önskar. Den sista meningen anger dock att han inte uppnått det målet.

Vad ip här gör är inte att tala om hur han har kommit till målet, utan varför han tycker som han gör, vilket uttrycks i form av förutsättningar i hans egen tankevärld. Om man skulle uttrycka detta i Abelsons symbolspråk kunde det kanske se ut som i figur 9.

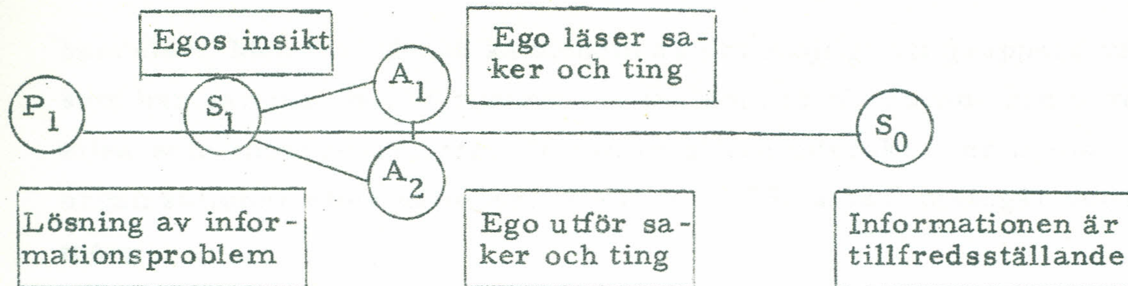


Figur 9. Representation av en plan

Figuren skall tydas på följande sätt:

Kedjan inom S_1 innefattar Egos erfarenhetsvärld. Värderingen av forskaren (P_2) skall vara en tilltro (S_5). Förutsättningarna ligger inom Ego själv (S_2, S_3, S_4). Eftersom detta är Egos kognitiva struktur, finns det ingen A-atom för handling och kedjan är ingen fullständigt egen plan, så vi har markerat den som tillhörig S_1 genom klammer.

Strukturen i S_1 medför att P_1 kommer till uttryck. En hämmsko för A_1 är dock S_6, som inbegriper ett P-S-par som uttrycker att A_1 inte kan genomföras på det sätt Ego skulle vilja, och därför når vi inte fram till S_0. Vad Ego vill är att någon (agent?) tar hand om problemet. Men finns det inget som Ego själv kan göra? Vi var intresserade av det och gick till materialet för att studera meningarna efter nr 41. Den information vi där fick ser ut så här i symbolform:



Han säger nämligen så här i fortsättningen:

"Det finns ju en annan lösning på informationsproblemet. Och det är ju helt enkelt att inse att antingen får jag ägna mig åt att läsa saker och ting eller också får jag ägna mig åt att göra saker och ting. Man kan inte göra båda delarna."

Ego har alltså två tänkbara strategier. Vilken han i realiteten ägnar sig åt framgår fortfarande inte. Vi kan här avsluta diskussionen genom att försiktigtvis konstatera att situationen för Ego är frustrerande.

Här har tagits upp ett besvärligt exempel på den nivå Abelson kallar plan. Det har medvetet valts för att konkretisera skillnaden mellan vad vi kan koda på ett syntaktiskt plan och vad som därmed går förlorat. Språkstruktur och tankestruktur går inte alltid hand i hand och det är den insikten som har varit avgörande för vad våra kodningsregler kan hjälpa till med och på vilken nivå vi sedan måste ha datorn till hjälp.

Den nivå Abelson (1973, s 315) kallar tema är en utvidgning av en plan, på så sätt att den involverar en annan actor. Han förklarar:

"... beliefs about social reality must include themes of cooperative and competitive interpersonal relationships. /.../ ... interdependence arises as a function of role in the other's plan, of attitude toward the purpose or activities of the other, and of ability to facilitate or interfere with the outcomes of the other. The conjunctions of these three variables gives each theme its particular character."

En actor kan vara målet för en handling av en annan actor, t ex så att handlingen är positivt eller negativt riktad mot en actor. Men en actor behöver inte vara direkt involverad utan endast som s k intressent. Dessutom har vi redan nämnt det förhållandet när en actor är agent till en annan. De känslor som olika actors har för varandra styr deras reaktioner, som kan vara handlingar eller kommentarer. De handlingar som olika actors utför med kännedom och åsikter om varandra kan utföras parallellt eller i ett beroendeförhållande, där den ena handlar åt den andra som agent. Agentens känslor för en actor bör i övervägande grad vara positiva, i annat fall blir han sin egen actor och tar egna initiativ som avviker mer eller mindre från den förras. På detta sätt resonerar Abelson sig fram till hur actors känslor för varandra kan pendla mellan kärlek och hat. Dessa och känslor däremellan, t ex "devotion", "admiration" är teman. För att kunna bygga upp en tematisk struktur måste man känna till vilka actor och agent är och vilka känslor som finns. Annars kan man inte bedöma en enskild

händelse. Men med denna kännedom är det möjligt att gruppera vilka actors som har samma belief system , vilka som är oberoende och beroende, vilka som kooper erar, etc. Teman är alltså interaktioner mellan individer, organisationer eller nationer. Abelson (1973, s 325) utvidgar detta resonemang:

"Themes capture what one might call 'temporary dynamics', relationships which last only as long as the durations of the respective plans. After the completion of the realization of a theme, a further realization of the same theme is possible only when the relationship between the actors remains the same."

När relationerna mellan actors ändras bildas nya teman. Ju längre utdraget händelseförloppet är i tiden, desto fler teman kan uppstå, s k temasekvenser eller scripts. Karakteristiskt för olika belief systems är oftast variationen av scripts. Vi skall ta ett exempel från Abelson (1973, s 320) för att klargöra skillnaden mellan tema och script.

Ett av de tolv olika teman som anges kallas dominans, som ju uttrycker beroendeförhållanden. Detta tema är det åttonde (T_8). Vi kan exemplifiera temaförändringarna i händelseförloppet (scriptet) revolution:

$$T_8\,(E;\ F) \longrightarrow T_9\,(E;\ F) \longrightarrow T_{12}\,(E;\ F) \longrightarrow T_7\,(E;\ F) \longrightarrow T_8\,(F;\ E)$$

Beteckningarna inom klammer är de tidigare kända symbolerna Actor (E) och Agent (F). Detta skript uttyds på följande sätt:

Utgångspunkten är att E dominerar F. Sedan inträder händelser, där rebelliska tendenser blommar upp (T_9), vilket leder till konflikt (T_{12}). T_7 säger att konflikten leder till seger för den underlydande och detta nya beroendeförhållande märks i F och E:s förändrade roll ifråga om dominans.

Abelsons nivåer representerar en hierarki, där teman och scripts befinner sig högst upp på pyramiden. För att strukturer på den nivån skall bli analyserbara måste först de föregående nivåerna ha urskilts. På de mer eller mindre konkreta nivåerna har vi kunnat göra en jämförelse med vårt sätt att representera strukturerna.

Abelson går inte in på hur hans system opererar genom de konkreta nivåerna för att resultera i en innehållsstruktur på de abstrakta, varför en jämförelse inte är möjlig. Den rekvisita som är nödvändig för att nå upp till de abstrakta strukturerna tycks dock i vårt fall finnas. Tematiska strukturer bygger på värderingar, vilket även för oss är grundläggande. Genom skalering av adjektiv och verb bestäms innehållet i de begrepp som bildar basstrukturen. En viktning av substantiv kommer att ske och denna procedur, som närmast skall rapporteras av B. Bierschenk, är inte genomförd i sin helhet. Så länge vi inte har innehållsbestämt våra grundenheter kan vi inte diskutera de olika temans relevans för vår ana-

lys och inte heller vilka kopplingar mellan eller inom nivåerna vi skulle behöva göra för att finna tematiska variationer. Detta medför att även scripts lämnas utanför i det här sammanhanget. Varken teman eller scripts kan tillföra våra kodningsregler något nytt. Frågan är om de överhuvudtaget är tillämpliga i vår analys. Relevansen är beroende av det innehåll vi vill nå fram till. Abelson (1973, s 334) definierar "innehåll" i sitt fall som:

"/an/ ... important source of variation between belief systems is in the assignment of particular actors or categories of actors to particular atoms, plans, themes, and scripts. This is essentially what we mean when we refer to 'content'."

Eftersom Abelsons syfte är att studera interaktioner mellan system, är det tveksamt hur långt vi kan gå. Den kunskap och värderingsstruktur som vårt material innehåller gäller inom ett system. Genom våra ip utsagor om olika actors kan vi få fram variationer i bedömningen av actors. Men det betyder inte att vi har definierat vilka personer som är actors och vilka som är agenter enligt Abelsons teori. (Jfr även fig 7.) Skillnaden mellan actor och agent är tämligen konkret inom ett politiskt system eller mellan system, men hur är det bland svenska forskare i pedagogik och psykologi? Är SÖ actor och forskare agenter i vissa fall? Det är onekligen en spännande frågeställning, men vi låter den tills vidare stå öppen.

7. SAMMANFATTNING

Denna rapport har redovisat ett försök med datorbaserad innehållsanalys, kallat ANACONDA (Analysis of concepts by data processing), som ingår som en metodstudie inom projektet Skolpedagogiska sökstrategier vid lärarhögskolan i Malmö.

Först har vår metod satts i ett internationellt sammanhang, på så sätt att den jämförts med internationella konventioner ifråga om behandlingen av text för vidare datoranalyser. I samband därmed redogjordes för de principer enligt vilka vi ämnar bygga upp begreppsregister, som skall bilda basen för vår analys (kap 2-4).

Därefter sattes vårt system i relation till en teori om språk- och tankestrukturer. Vi har visat på vilket sätt den modell, som innehåller en symbolisk representation av begrepp och relationer, och vars teoretiska antaganden vi ansluter oss till, är möjlig att omsätta till ett system av sekventiella sifferkombinationer (kap 5).

Till sist anknöt vi till en teori om "belief systems", där vi kunde förankra våra kodningsregler i praktiken, dvs där våra kodningsregler kunde tillämpas i en hierarkisk struktur. I detta avsnitt försökte vi komma fram till på vilken hierarkisk nivå i systemet våra kodningsregler är till hjälp vid analys av tankestrukturer och i vad mån andra nivåer i detta system är relevanta för oss (kap 6).

En summering av de ovan nämnda jämförelserna görs nedan. Avsikten med denna relativt utförliga sammanfattning är att ge läsaren möjlighet att informera sig om innehållet i rapporten utan att behöva studera varje kapitel ingående. Den som är intresserad av de mera tekniska aspekterna rekommenderas dock att följa jämförelsen kapitel för kapitel.

7.1 Konventionell behandling av text och uppbyggnad av lexikonbas

En innehållsanalys baseras alltid på ett lexikon, men förutsättningarna för detta kan variera. Vi har jämfört vår analys med General Inquirer (GI), som är ett av de mest kända textanalyssystemen inom samhällsvetenskapen. (Se t ex Stone et al., 1966.) Denna jämförelse har gett oss följande väsentliga skillnader att redovisa.

Texterna som behandlas i GI-systemet är vanligen skrivna dokument, medan vi har talspråkstext, vilket medför att vi måste tillgripa särskild rensning av texten innan vi kan koda den.

Analysenheten i GI är det enskilda ordet och man stansar in hela meningar utan segmentering. Vår analysenhet är meningen "sentence", där delarna inte är ord utan begrepp, som kan bestå av flera ord. Genom att

vi kodar meningar kan vi också koda det övergripande temat, vilket GI inte gör.

GI redovisar problem med idiom, vilket för oss inte innebär problem, eftersom dessa kodas som ett begrepp, oavsett hur många ord det innehåller.

Användare av GI får ingen syntaktisk information. Vi kodar denna i form av AaO-relationer.

En internationell konferens om datorbaserad innehållsanalys inom samhällsvetenskapen, som hölls i Pisa 1974,har kommit fram till att en standardisering måste ske ifråga om bl a inmatning och identifikation av items. Innan denna konferens hölls hade vi utarbetat principerna för vår input. Vi har kunnat konstatera att vi följer dessa s k Pisa-konventioner, vilket är viktigt om vi vill presentera vårt system internationellt.

Pisa-konferensen mynnade också ut i några rekommendationer för forskare och systemutvecklare när det gäller lexikonbasen. GI har problem med att särskilja typ av ord och fall av samma ord. Ett ord (i vårt fall begrepp) i lexikon skall kategoriseras genom deskriptorer, som särskiljer det från andra ord (begrepp). De problem som diskuteras gäller särskilt sammansättningar och suffixbildningar, där inte alla språk är uppbyggda som engelskan, vilket GI grundar sig på. Vi har visat att det är möjligt att bygga upp lexikon på svensk bas och hävdar att det är nödvändigt för varje språk att utbilda sina egna regler för registrering av items i lexikon. Vi använder bl a ett trunkeringsförfarande, som medger en stor flexibilitet när det gäller att återvinna begrepp, sammansatta eller enkla. Jämfört med GI tycks vårt sätt att arbeta med begrepp i stället för ord betydligt förenkla denna registreringsprocedur.

Vi har betonat att begreppen i vår analys i motsats till GI är kontextbundna och denna bundenhet framträder redan på kodningsstadiet (AaO-paradigmet). Vi avser inte utveckla program där datorn skall "förstå" texten. Det innebär att vår kodning av text är en tidsödande procedur, men i gengäld ger den större möjlighet än GI att snabbt ge oss svar på frågor, vilket GI inte har möjlighet till. T ex har vi redan på kodningsstadiet differentierat mellan olika typer av meningar. Att få fram sådana differenser i efterhand är ett omfattande programmeringsarbete.

Sammanfattningsvis kan sägas att de problemområden som presenterats till konferensen i Pisa i hög grad faller inom disciplinen sociologi, vilket kan förklara det ringa intresset för språkstruktur eller psykologiska teorier. Dessa förankringar hos oss har medfört det relativt stora avståndet från de GI-baserade textanalyserna.

7. 2 Kodningsregler enligt en teori om tankestrukturer

Vi ansluter oss till Schanks (1973) teori om "conceptual dependencies". Enligt den rör sig språkanalys på två nivåer, en meningsnivå ("sentential level") och en begreppsnivå ("conceptual level"). Det finns egentligen bara två slag av begrepp, nämligen beroende och oberoende. Detta förhållande har vi uttryckt i vårt kodningssystem genom tvåsiffriga koder.

Schank presenterar 14 grundregler som gäller de båda nivåerna. Dessa regler har jämförts med ANACONDA-kodningen, där vi kommit fram till följande resultat.

Vi har i vår framställning urskilt flera slag av beroenden, dels ett vertikalt, dvs mellan beroende och oberoende begrepp, dels mellan begrepp på ett syntaktiskt horisontellt plan. Vi diskuterar också beroende och oberoende relationer, dvs satsers förhållande till varandra. Det första slaget av beroende avser Schanks regler 1-5. Oberoende begrepp är ,sådana, som är huvudord i ett komplex, bestående av t ex attributiva bestämningar till detta begrepp. Huvudbegreppet har en kodsiffra som slutar på 0, medan beroende begrepps slutsiffra är 1, 2 eller 3. Reglerna 1-5 är representerade i vårt kodningssystem utan inskränkningar.

Reglerna 6-9 gäller s k conceptuella kasus och avser ett horisontellt beroende, såtillvida att man med utgångspunkt i verbets betydelse specificerar hur många oberoende begrepp som måste finnas på meningsnivån. Dessa kasusförhållanden framträder även hos oss, men vi har hittills inte benämnt koderna enligt kasus. De fyra kasus är "objective", "recipient", "instrumental" och "directive". De motsvaras hos oss av respektive objekt (direkt), objekt (indirekt, eller mål), sättsadverbial och rumsadverbial. Det enda fall där vi kan säga att vår kodning egentligen skiljer sig från kasusdistinktioner är i fallet "directive." Vi har nämligen inte ännu skilt på riktning och befintlighet i denna kod.

Angående beroendets uttryck i kodsiffrorna kan sägas att kasus är oberoende begrepp (substantiv) som i sin interna beroendestruktur inte skiljer sig åt. Därför borde de inte ha en beroendesiffra, som ansluts till verbet (andra kodsiffran). Eftersom vi inte från början utgått från kasusteori ("conceptual cases") har vi dock i två fall uttryckt ett sådant beroende, nämligen i adverbialen. Där har vi tagit hänsyn till det traditionella sättet att analysera satser och dessa adverbialbegrepp är lätta att känna igen vid kodningen. Det är således en praktisk åtgärd och har inget med avvikelse från Schanks teori att göra. Men vi kommer förmodligen att ändra våra koder så att de på ett bättre sätt ansluter till teorin, särskilt i fallet "directive".

Teorin om nödvandiga delar tar vi även hänsyn till på annat sätt. Eftersom språk är ekonomiskt i förhållande till tankar bakom yttranden, finns inte alla nödvändiga delar med i en talare-lyssnare-situation. Den nödvändiga begreppsapparaten finns hos lyssnaren och en syntaktiskt ofullständig mening blir ändå förstådd. Men denna förståelse har inte datorn, så därför supplerar vi in de delar som saknas. Det gäller för oss att koda fullständiga komplex och fullständiga AaO-paradigm. Paradigmet innebär agent, verb och objekt (ett eller två). När så verbet i stället för objekt har plats som mål suppleras detta begrepp in. I paradigmet ingår däremot inte instrument som nödvändigt, om det inte finns explicit. En väsentlig skillnad mellan Schanks teori och vår praktiska kodning i anslutning till den gäller hur långt vi skall gå i representation av conceptuella regler (tankestrukturer). Hos Schank är kasus "instrument" en nödvändig del inom verbets betydelse, så att t ex verbet ′äta′ måste betyda ungefär ′med bestick′ som instrument. Som ovan sagts kodar vi inga instrument som inte nämnts i texten. På samma sätt blir även rumsadverbial utanför paradigmet och kodas inte, även om man måste förutsätta att en handling ägde rum någonstans. Men Schank säger också själv att det finns vissa begrepp som är så kända av lyssnaren (t ex ′äta′) att det är ointressant att specificera instrumentet. Vi tänker inte på det medvetet.

Schanks regler 11 (a, b) - 14 gäller vad han kallar relationer, dvs satsers förhållande till varandra inom en mening. Vi har kunnat visa att, även om hans modell i form av diagram är symboliskt komplicerad , vi har möjlighet att uttrycka relationer genom vårt siffersystem i kombination med övergripande kodning av ett s k satstema.

Även här avspeglas en skillnad i sättet att representera verbet. Schank symboliserar explicit det resultat som ligger i vissa verb, t ex ′döda′, som han kallar pseudo-state-verbs. Vi stannar vid att koda verbet och anger inte ett eventuellt resultat. Det medför att regel 11 b för oss inte är relevant.

Relationer framträder genom en satsinledare. Dessutom kodas satsberoende med koder för orsak, avsikt, osv. Ett loop-system gör det möjligt att urskilja huvudsats och bisats eller vilken av satserna som kommer tidsmässigt före den andra. Det är detta som Schank kallar kausalitet. Det finns även andra conceptuella relationer, nämligen tids- och platsrelationer. Tid som ett enda begrepp eller som bestämning i form av bisats ses som modifierare till en hel sats och det finns alltså ingen kausalitet. Det mest praktiska sättet att ange modifiering till en hel sats har vi tyckt vara att ge dessa begrepp en beroendekod till verbet, eftersom det är verbet

som är det väsentliga i satsen. Att ingen kausalitet finns anges genom att satsen inte får satstemakod, som de övriga. Reglerna 12-14 återspeglas alltså hos oss.

Sammanfattningsvis kan vi anföra följande. Den teori som Schank presenterar och som vi ansluter oss till har fått sitt uttryck i den kodning av text som har utvecklats inom ANACONDA-systemet. En tveksam punkt finns dock, nämligen i vad mån det är nödvändigt att explicit ansluta sig till kasus-begreppen. Möjligen vore det av värde att se över adverbialkoderna och splittra dessa på samma sätt som de övriga koder som innehåller substantiv, dvs oberoende begrepp.

7.3 Kodningsreglers tillämpning i en hierarkisk struktur

En kodning av en text enligt en språkteori behöver ofta modifieras i enlighet med en teori om den "värld" som dessa data är en representation av. En sådan modifiering av Schanks teori har vi funnit hos Abelson (1973) i ett arbete om "belief systems", vilket därför har diskuterats i jämförande syfte.

Det för vår del väsentliga ställningstagandet hos Abelson är att han är mer orienterad mot relationer mellan beroenden än inom beroenden. Detta får bl a till följd att han inte intresserar sig för verbs implicita resultat utan den aktivitet som verbet uttrycker genom riktningen i handlingen. Där har vi således samma utgångspunkter.

En annan punkt på begreppsplanet, där vi får stöd av Abelson är att han i första hand arbetar med huvudbegrepp. Bestämningar av olika slag är inte nödvändiga delar. Det är när bestämning finns utan huvudord som supplering sker hos oss och inte tvärtom.

Abelson utgår ifrån att det finns sex nivåer i ett belief system, nämligen element, atomer, molekyler, planer, teman och scripts. Element motsvaras av våra begrepp, atomer av satser. Molekyler är länkar av satser som förs ihop, om satsernas verb uttrycker tanke (syfte), handling eller tillstånd. Kedjor eller nätverk av sådana molekyler bildar en plan, som innefattar syfte (avsikt eller mål), handlingar eller tillstånd som direkt eller sidoordnade för till målet, samt det slutliga tillstånd som varit syftet. Dessa nivåer kan betecknas som konkreta och har kunnat jämföras med våra kodningsregler och deras kapacitet ifråga om återvinning. De övriga två nivåerna, de abstrakta, innebär strukturer i värderingar och attityder hos personer, organisationer och nationer och har inte kunnat ge våra kodningsregler något tillskott. Dessa nivåer kan endast diskuteras när en innehållsbestämning har gjorts av våra begrepp. I vad mån dessa översta nivåer i hierarkin är applicerbara på vårt material och re-

levanta för det belief system vi analyserar har ifrågasatts, främst med tanke på att vi rör oss inom ett system och inte mellan system.

Det väsentliga i diskussionen om de fyra konkreta nivåerna är än en gång att påpeka att vår största analysenhet är den syntaktiska meningen. För oss kan således varje typ av atom vara en mening, antingen den uttrycker tanke, handling eller tillstånd. Ett tillstånd uttrycks via vår kodning genom att verbet 'vara' m fl särskiljs med annan kod från övriga verb. Utan att ytterligare kategorisera verben vid kodningen kan vi visa att vi kan återvinna sådana fullständiga molekyler som Abelson antar.

En speciell distinktion gör Abelson, nämligen den mellan "actor" och "agent". Det tycks särskilt lämpligt i det politiska spel han diskuterar utifrån. Actor är överordnad agent och den senare ses som beroende. Vi har inte utan vidare kunnat anta den distinktionen, eftersom vår analys inte har till syfte att spegla sådana beroenden mellan personer. Men vi har kunnat visa att det går att på ett syntaktiskt plan komma fram till den genom huvudsats- och bisats-tekniken, förutsatt att detta beroende är explicit på den syntaktiska nivån.

Det har varit särskilt väsentligt att framhålla våra sats-tema-koders betydelse för en analys av strukturer på olika nivåer. Genom att vi specificerar en sökning i materialet till att gälla meningar med kod för t ex avsikt, kan vi återvinna vad Abelson menar med "purpose." Vi kan sedan fortsätta bakåt eller framåt och få fram handlingar och tillstånd, så att vi kommer fram till nivån plan . Planer är utvidgade molekyler, dvs en serie kodade meningar av varierande längd. Våra kodningsregler sträcker sig fram till molekyl-planet. En sökning med hjälp av satstema-koder för oss fram ytterligare ett steg, men därefter tar de abstrakta nivåerna vid.

Sammanfattningsvis kan sägas följande. Den teori om belief systems som Abelson presenterar har som grund Schanks begreppsstrukturer. Abelson går ett steg ytterligare genom sin analys av beroenden mellan relationer. Vår grund är densamma och omfattar även de inskränkningar av Schanks modell som Abelson gör. Därutöver har vi kodningsregler som täcker beroenden över meningsplanet upp till en struktur av den nivå som kallas plan. För att systematiskt kunna återvinna planer bör vi emellertid se över verben, på så sätt att vi bestämmer vilka som innebär tillstånd (t ex även verbet 'ha'). Detta kan bli möjligt genom resultatet av den skaleringsprocedur som genomförts och som skall redovisas inom kort.

8. REFERENSER

Abelson, R.P. The structure of belief systems. I: Schank, R.C. & Colby, K.M. Computer models of thought and language. San Francisco: Freeman, 1973. Ss 287-339.

Bierschenk, B. Datorbaserad litteratursökning. Lund: Studentlitteratur, 1973.

Bierschenk, B. A computer-based content analysis of interview data: Some problems in the construction and application of coding rules. Didakometry (Malmö, Sweden: School of Education), No. 45, 1974. (a)

Bierschenk, B. Perception, strukturering och precisering av pedagogiska och psykologiska forskningsproblem på pedagogiska institutioner i Sverige. Pedagogisk-psykologiska problem (Malmö: Lärarhögskolan), Nr 254, 1974. (b)

Bierschenk, I. Konstruktion av ett regelsystem för en datorbaserad innehållsanalys av intervjutext: Preliminärmanual och några utprövningsresultat. Testkonstruktion och testdata (Malmö: Lärarhögskolan), Nr 25, 1974.

Bierschenk, I. Skalering av adjektiv och verb med hjälp av semantiska differentialer: Urval av begrepp och praktiska förberedelser. Stencil (Malmö: Lärarhögskolan), April, 1975.

Coxon, A.P.M. & Chalmers, A.D. Content analysis as a social science resource. Annual progress report. Stencil (Edinburgh: Social Science Research Council), 1973.

Deichsel, A. Social structure of symbolic worlds. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 69-76.

Fonnes, I. TEXT - a general program package for text processing. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 77-83.

Garbett, S. The application of content analysis to a research project entitled: Televising parliamentarians and the mass media interact and what is and could be the role of television in this context. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 85-98.

Himmelstrand, U., et al. The nigerian civil war in international press. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 19-48.

Holzschenck, K. System for action analysis (abstract). Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 287-289.

ISSC/CISS. Workshop on content analysis in the social sciences. Pisa: CNUCE - Istituto del CNR, September, 1974.

Kops, M. The identification of person conceptions by the content analysis of marital advertisements (abstract). Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 291-294.

Kächele, H. et al. Changes of speech content during a psychoanalytic process. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 99-110.

Maranda, P. Characterization of characters' characteristics: Is a work of art. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 111-123.

Maranda, P. & Maranda, E.K. Myth as a cognitive map: A sketch of the okanagan automaton. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 125-153.

Mochmann, E. Z.A.R.: A content analytical approach to indicator retrieval. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 285-286.

Mochmann, E. Cont·nt analysis techniques retrieved from approaches to information retrieval. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 297-298.

Mohler, P. Relation between individual and collective power as it appears in German boy's graduation compositions from 1917 to 1971. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 299-301.

Namenwirth, J.Z. Change within or of the system: An example from the history of American values. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 155-178.

Schank, R.C. Identification of conceptualizations underlying natural language. I: Schank, R.C. & Colby, K.M. Computer models of thought and language. San Francisco: Freeman, 1973. Ss 187-247.

Siune, K. Structure and content of the 1971 election campaign in Danish radio and television. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 179-200.

Spence, D. PL/1 programs for content analysis. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 201-232.

Stone, P., Dunphy, D., Smith, M. & Ogilvie, D. et al. The general inquirer. Cambridge: M.I.T. Press, 1966.

Stone, P. Report on the workshop on content analysis in the social sciences Pisa, 1974. Soc. sci. inform., 1975, 14 (1), 107-111. (a)

Stone, P. Standards for computer-aided content analysis: The Pisa conventions and recommendations. Soc. sci. inform., 1975, 14 (1), 127-137. (b)

Tiemann, R. Developing variables for a dictionary of ingroup-outgroup analysis. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 303-305.

Transgaard, H. Brief presentation of current research. Paper presenterat vid ISSC/CISS workshop. Pisa: CNUCE, 1974. Ss 307-308.

Trappes-Lomax, H.R. A computer based system for content analysis. A review of the Edinburgh new tagger version of the general inquirer. Stencil (Edinburgh: Department of Sociology), 1974.

9. BILAGA

9.1 Termförklaringar

9.1 Termförklaringar

final bisats	uttrycker en avsikt av typen ..., för att ...
idiom	är uttryck som består av flera ord, som tagna var för sig inte betyder detsamma som inom flerordsgruppen. Man brukar säga att idiom är stelnade uttryck, vilket bl a innebär att ett substantiv som ingår i uttrycket inte ändrar form i vilket sammanhang det än förekommer: 'bära hand på', 'hålla hus'. Förutom dessa uttryck finns andra vanliga flerordsuttryck, som t ex 'i dag', 'för alltid', som också räknas till de idiomatiska.
kasus	fall, verklig eller möjlig eller tänkt händelse eller omständighet; benämning på böjningsform som anger ett nomens (substantiv) eller pronomens förhållande till annat ord eller ordgrupp i satsen. Kasus bestäms genom verbets betydelse. Vissa verb "fordrar" fler kasus än andra. Det kasus (eller fall) ett nomen får i en viss sats kan ibland anges genom prepositioner i svenskan (såsom åt, för, etc). Kasus-böjning i t ex ackusativ och dativ finns i tyskan.
kausal bisats	uttrycker en orsak av typen 'Eftersom ...', 'Av att ...'
koncessiv bisats	uttrycker ett medgivande av typen 'Trots att ..., 'Även om ...'
konditional bisats	uttrycker ett villkor av typen 'Om ..., (så)'
konsekutiv bisats	uttrycker en följd av typen '..., så att ...'
kontext	sammanhang (vari ett ord eller uttryck ingår), omgivande text
kontextbunden	med hänsyn tagen till omgivande text
kontextfri	utan hänsyn tagen till omgivande text
kopula	förbindelseled, verbform som förbinder subjekt och predikat i en sats. Det vanligaste kopulaverbet är 'vara' ('är'), dessutom 'bliva', 'heta', 'kallas' m fl liknande.
modal, modalitet	(se modus)
modus	det sätt på vilket något framträder som varande eller handlande; formkategori som anger den talandes (eller skrivandes) uppfattning av det utsagda. Indikativ, konjunktiv och imperativ är modus. Indikativ uttrycker något som faktiskt sker. Imperativ uttrycker uppmaning. Konjunktiv uttrycker inget faktiskt skeende, utan den talandes (eller skrivandes) önskan, etc. När vi talar om modalitet, menar vi vanligen den betydelse som ligger i konjunktivens funktion, även om konjunktiv som verbform sällan förekommer numera i svenskan. Modalitet kan uttryckas förutom genom verbformen också genom modifierande adverb (kanske, möjligen, etc) eller genom modala hjälpverb (vill, bör, måste, etc).

morfologisk	av 'morfologi' som betyder formlära. Den minsta betydelsebärande enheten kallas morfem. Morfemen kan vara lexikaliska(lexem)eller grammatiska (ändelser, prefix, suffix).
predikativ	det satsled som följer efter ett kopulaverb; i traditionell satsanalys även kallat predikatsfyllnad. Predikativa påståenden är sådana påståendesatser (dvs som inte uttrycker fråga, uppmaning eller liknande) som konstrueras med kopulaverb som länk mellan subjektet och övriga satsled.
semantisk	av 'semantik' som betyder betydelselära.
syntaktisk	(se syntax)
syntax	satslära, del av grammatiken som behandlar ordens förbindelser till högre enheter. Syntaktiska regler säger t ex i vilken ordning morfem binds ihop till ord, ord till fraser och fraser till satser och meningar. Om enheterna inte följer en viss ordning (struktur) är en sats inte syntaktiskt korrekt.
tempus	tidsform hos verb. Presens, futurum, etc är tempusbeteckningar. Verb har en tidsaspekt och en modusaspekt (se modus).

Department of
Educational and
Psychological Research
School of Education
Malmö, Sweden

Abstract card

Bierschenk, I. Datorbaserad innehållsanalys: Teoretiska och praktiska överväganden. /Computer-based content analysis: Theoretical and practical considerations./ Pedagogisk-psykologiska problem (Malmö, Sweden: School of Education), Nr 283, 1975.

This report describes a Swedish contribution to computer-based content analysis research. The method is compared to international conventions regarding the treatment of text for computer-based analysis. An account is given of how a dictionary can be built up on the basis of the Swedish language. The coding of the text is based on a theory about thought structures and the report describes the way in which this theory can be converted into a system for coding concepts and relations. The system of rules is then put in relation to a theory about the simulation of cognitive and emotional structures. The capacity of the system of rules for analysis on different hierarchical levels is presented.

Indexed:
1. Content analysis
2. Computer-based language analysis
3. Concept formation

Reference card

Bierschenk, I. Datorbaserad innehållsanalys: Teoretiska och praktiska överväganden. /Computer-based content analysis: Theoretical and practical considerations./ Pedagogisk-psykologiska problem (Malmö, Sweden: School of Education), Nr 283, 1975.

ISSN 0346-5004